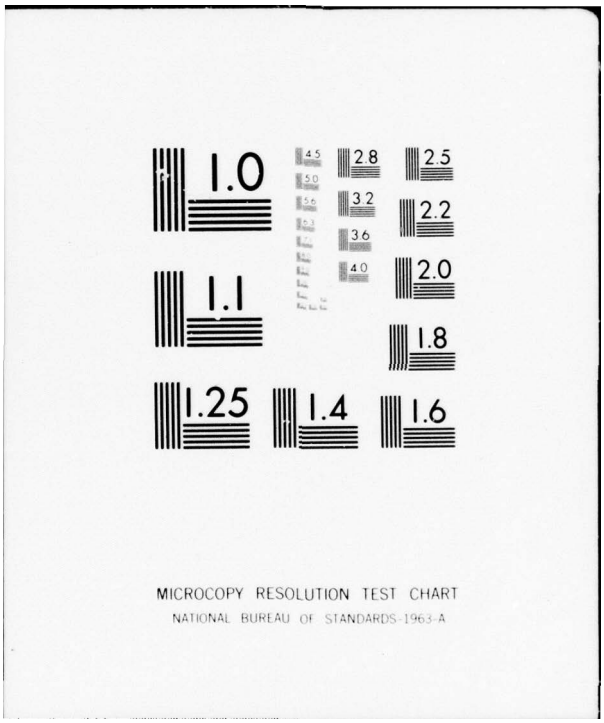
AD-A037 483 AIR FORCE FLIGHT DYNAMICS LAB WRIGHT-PATTERSON AFB OHIO F/G 12/1
STATISTICAL MEASURES, PROBABILITY DENSITIES, AND MATHEMATICAL M--ETC(U)
OCT 76 R G MERKLE

UNCLASSIFIED AFFDL-TR-76-83 NL

1 OF 2
AD A037483


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD A037 483

AFFDL-TR-76-83

STATISTICAL MEASURES, PROBABILITY DENSITIES, AND MATHEMATICAL MODELS FOR STOCHASTIC MEASUREMENTS

FIELD TEST AND EVALUATION
STRUCTURAL MECHANICS DIVISION

OCTOBER 1976

TECHNICAL REPORT AFFDL-TR-76-83
FINAL REPORT FOR PERIOD NOVEMBER 1973 - JULY 1976

AD No. DDC FILE COPY

DDC MAR 30 1977

Approved for public release; distribution unlimited

AIR FORCE FLIGHT DYNAMICS LABORATORY
AIR FORCE WRIGHT AERONAUTICAL LABORATORIES
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

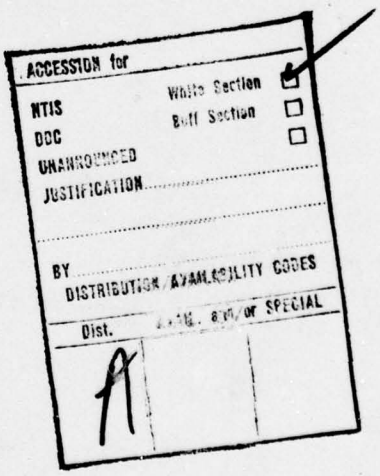
This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

Robert G. Merkle
ROBERT G. MERKLE
Project Engineer

FOR THE COMMANDER

Howard L Farmer
HOWARD L. FARMER, Col, USAF
Chief, Structural Mechanics Division



Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

AIR FORCE - 23 MARCH 77 - 400

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFFDL-TR-76-83	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): STATISTICAL MEASURES, PROBABILITY DENSITIES, AND MATHEMATICAL MODELS FOR STOCHASTIC MEASUREMENTS	5. TYPE OF REPORT & PERIOD COVERED: Final Report November 1973 - July 1976
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Robert G. Merkle	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Field Test and Evaluation (FBG), Air Force Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio 45433	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Project No. 1472, Task No. 147202, Work Unit No. 14720204
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio 45433	12. REPORT DATE: October 1976
	13. NUMBER OF PAGES: 121
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

012 070

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistics
Probability Density Functions
Correlation
Factor Analysis
General Linear Hypothesis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Statistical measures, probability densities, and mathematical modeling techniques useful in the analysis of stochastic measurements are summarized. Univariate measures are given for average, dispersion, skewness, and kurtosis. Probability Densities include: Normal, Student t, Cauchy, Gamma, Exponential, Chi-square, F, Rayleigh, Maxwell, Log-normal, Beta, Uniform, and Arc-sine. Measures of interdependence between two variables include simple correlation, autocorrelation, cross-correlation, rank correlation, point biserial → next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

6 pg

cont →

20. ABSTRACT (Cont'd)

correlation, tetrachoric correlation, and coefficients of contingency. Measures of interdependence among several variables include multiple correlation, marginal correlation, conditional correlation, canonical correlation, and auto and cross-correlation for ensembles of measurements. Mathematical modeling techniques include factor analysis and both regression and analysis of variance formulated as the general linear hypothesis model.

FOREWORD

This report was prepared in the Structural Mechanics Division, Field Test and Evaluation Branch, Air Force Flight Dynamics Laboratory (AFFDL/FBG), Wright-Patterson Air Force Base, Ohio. The work was done under Project No. 1472 "Dynamic Measurement and Analysis Technology for Flight Vehicles", Task 147202 "Dynamic Data Analysis for Flight Vehicles", Work Unit 14720204 "Statistical Reduction of Dynamic Data." The research covered the period November 1973 through July 1976. Mr. Robert G. Merkle was the responsible engineer and author of this report.

The probability density plots that appear as Figures 4 through 11 were programmed and produced by Mary Folz. Figure 13, computation of aurocorrelation values was taken from reports prepared for the government by J. S. Bendat, A. G. Piersol, and L. D. Enochson. Other figures were produced by Mary Jo Bornhorst. Skewness and Kurtosis coefficients for the Rayleigh and Maxwell Distributions were computed by H. L. Harter. The derivations in Appendices A, B, C, and D were pointed out by Robert J. Wherry, Psychology Dept, Ohio State University. Typing was done by Timothy Ketzel and Dorothy C. Young. Drafting of mathematical expressions was done by James Sommerville and John Skinner.

TABLE OF CONTENTS

PAGE BLANK-NOT FILMED

TABLE OF CONTENTS (Cont'd)

LIST OF FIGURES

SECTION I

INTRODUCTION

Measurements taken from systems responding to dynamic excitations generally have an unpredictable random element characteristic of one or more of the excitation forces. Because of this randomness, multiple observations are necessary. Modern automated multichannel instrumentation systems are capable of sensing and recording an enormous volume of these excitation and response measurements from numerous locations and test conditions. This report summarizes a number of definitions and analysis methods that are especially useful in the statistical treatment of such voluminous data.

1. STATISTICAL MEASURES FOR SINGLE VARIABLES

An average value of some kind is generally considered to be the most important statistic since it is the single value considered best to represent an entire set of observations. Six different measures of average value are defined: the mode, median, arithmetic mean, quadratic mean, harmonic mean, and the geometric mean. The choice of which to use depends on the particular application and some guidelines are given for making this selection.

The dispersion of a set of observations is generally the second most important statistic since it is the single value best representing the degree of scatter in the data. Five different measures of dispersion are defined: the range, mean deviation, standard deviation, variance, and the coefficient of variation. As with the average, the choice of which to use depends on the particular application and some guidelines for this choice are given.

PRECEDING PAGE BLANK-NOT FILMED

The symmetry and relative concentration about the mean are two other statistics important in describing the distribution of a set of observations. The coefficients of skewness and kurtosis are defined to measure these two attributes. These two statistics are often useful in determining what mathematical probability density functions are consistent with the measured data.

2. PROBABILITY DENSITY FUNCTIONS FOR SINGLE VARIABLES

Univariate probability density functions are mathematical functions of one variable which plot as continuous single valued curves lying on or above the horizontal coordinate axis with a finite area between that axis and the curve. The function is multiplied by the reciprocal of this area as a normalizing constant, so that the fractional area above any axis interval represents the proportion of the random variable observations having values within that interval.

The probability density curve may be unbounded on both sides extending from minus infinity to plus infinity along the horizontal axis; it may be bounded on one side as in the case of all positive numbers extending from zero to plus infinity; or it may be bounded on both sides as in the case of proportions extending from zero to plus one. Using this classification system important probability density functions will be defined, with the mean, standard deviation, skewness, and kurtosis of each given in terms of the constant parameters appearing in the defining equation.

The Normal, Student t, and Cauchy Probability Density Functions are the unbounded types defined. The normal is important not only as a parent population for many measurements, but also as the sampling distribution for the mean. Under fairly general conditions, the sampling distribution of mean values computed from numerous independent samples approaches the normal distribution as the sample size increases even if the parent population of the sample is not normal. This is the reason for the importance of the normal distribution in statistics. The Student t distribution is the sampling distribution of the difference between population and sample means divided by the ratio of the sample standard deviation to the square root of the sample size. It approaches the normal as the sample size increases. The Cauchy distribution arises from the quotients of two independently distributed normal observations having zero means. If both normal variates have unit variance the resulting Cauchy distribution is the same as a Student t for a sample size of two. The Cauchy density is of interest because it has infinite variance. Consequently, if two independent normally distributed observations with zero means are divided in the course of data processing operations, nothing at all is gained by increasing the sample size of such quotients.

The Gamma, F, Rayleigh, Maxwell, and Log-normal Probability Density Functions are the types defined for all positive numbers. The Gamma is important not only as a parent population for many measurements but also because of two important special cases; the

exponential and chi-square distributions. The time intervals between independent randomly occurring events are exponentially distributed. The ratios of sample to population variance for independent normal samples have a chi-square over degrees of freedom distribution. The ratios of sample variances for pairs of independent normal samples have an F distribution. If both rectangular coordinates for observations in a plane have normal distributions, then the radii in polar coordinates have a Rayleigh distribution. If all three rectangular coordinates for observations in space have normal distributions, then the radii in spherical coordinates have a Maxwell distribution. If the logarithms of measurements are normally distributed, then the measurements themselves are said to have a log-normal distribution.

The Beta Probability Density Function is one of prime importance for variables bounded on both sides, that is having finite upper and lower limits. Two special cases are of particular interest. A constant probability throughout the range of the variable forms a uniform distribution. The ordinates of a sine wave follow an arc-sine distribution. If both rectangular coordinates for observations in a plane have normal distributions, then the angles in polar coordinates have a uniform distribution and the sines of those angles have an arc sine distribution.

3. SIGNIFICANCE TESTS FOR PAIRS OF MEANS AND VARIANCES

The mean and variance are defined, respectively, as measures of average value and dispersion in Section *II*. *The t and F probability* densities are given, respectively, as distributions for sample means and sample variance ratios in Section III. In Section IV these two distributions are used in statistical tests for significant differences in the means or variances of two samples.

If the means of two samples differ significantly, then those differences in location or test condition under which the two samples were recorded are variables affecting the measured values. This is the normal experimental situation, the sample measurements having been made precisely to determine if such differences in the sample conditions are associated with differences in expected magnitudes.

If the variances of two samples differ significantly, then those differences in location or test condition under which the two samples were recorded are variables affecting the measured values. However, this necessary conclusion is often missed if attention is focused only on average values. Significantly different variances about similar means imply greater dispersion in the one sample than the other and the reason for this effect is generally of considerable interest.

4. STATISTICAL MEASURES OF INTERDEPENDENCE AMONG VARIABLES

The coefficient of correlation is the statistic measuring the degree of interdependence between two variables. A perfect correlation

of plus one between two variables x and y implies that they may differ only in the reference point and scaling unit so that $y = a + bx$ or $y = \bar{y} + b\ (x - \bar{x})$ with $b > 0$. A perfect correlation of minus one implies that $y = a - bx$ or $y = \bar{y} - b(x - \bar{x})$ with $b > 0$. Zero correlation implies no relationship at all. Intermediate positive and negative values, of course, imply imperfect correlation in which an error term e varying randomly with each measurement appears in the relationship, $y = a + bx + e$.

Eight types of bivariate correlations are defined to accommodate simultaneous or sequential types of quantitative data and dichotomous or multichotomous types of qualitative classification criteria.

a. Simple correlation is computed from a set of paired measurements of two different quantitative variables.

b. Auto correlation (or serial correlation) is computed from a single set of sequential measurements with each value representing an observation of the first variable and the corresponding observation of the second variable given by the value a fixed number of steps later in the sequence. This is useful in finding any periodicities in a sequence of measurements.

c. Cross-correlation is computed from a dual set of sequential measurements representing two different quantitative variables with the second variable of the paired observations displaced a fixed number of steps later in its sequence than the first. This is useful in excitation-response relationships.

d. Rank Correlation is computed from a set of paired ranks obtained for each bivariate observation from their positions in the ordered sequence of values for each variable. This is useful when one or both of the variables cannot be measured directly but can be ranked according to size.

e. Point-Biserial Correlation is computed from a set of paired observations one of which is the value of a quantitative variable and the other is a zero or one value of a qualitative variable expressing the absence or presence of a given attribute.

f. Tetrachoric Correlation is computed from a set of paired observations both of which are zero or one values expressing the absence or presence of different attributes.

g. Coefficient of Contingency and Correlation of Attributes are both computed from contingency tables showing the number of observations in each cell of a matrix in which the number of rows and the number of columns represent the numbers of classification categories for the two variables. In both cases identical row distributions in all columns and identical column distributions in all rows imply zero correlation. For square matrices all observations on the diagonal imply perfect correlation and the correlation of attributes is one. For nonsquare matrices there is no unique diagonal and the coefficient of contingency has a maximum value less than one. Whatever the type, even perfect correlations cannot alone indicate a cause-effect relation. Variations in either may be caused by changes in the other, or variations in both may be caused by changes in some third variable.

Six types of multiple correlation are defined, the first four computed from matrix arrays of bivariate correlations and the last two computed from ensembles of sequenced observations.

a. Multiple Correlation measures the degree of relationship between one dependent variable and an entire set of independent variables all taken together.

b. Marginal Correlation measures the degree of relationship between one dependent variable and a subset of the independent variables with all remaining independent variables simply ignored.

c. Conditional or Partial Correlation measures the degree of relationship between one dependent variable and a subset of the independent variables after statistically adjusting for the effects of all the remaining independent variables.

d. Canonical Correlation measures the degrees of relationship between a set of dependent variables and a set of independent variables. Marginal and conditional canonical correlations could also be defined as above by either ignoring or statistically adjusting for a subset of the independent variables.

e. Auto correlation measures the degree of relationship between measurements taken at any pair of sequence points from an ensemble of sequenced data records. Varying the pair of sequenced points leads to a matrix of correlations.

f. Cross-correlation measures the degree of relationship between measurements taken at one fixed sequence point from one ensemble and at another fixed sequence point from another ensemble. Varying this pair of sequence points leads to a correlation matrix. The two ensembles are the set of excitation or stimulus points and the set of response points.

5. FACTOR ANALYSIS OF MULTIPLE VARIABLES

Large numbers of observations of a single variable are reduced to a few statistics describing average value, dispersion, skewness, and kurtosis. Large numbers of interrelated variables may also be statistically reduced to a relatively few independent factors describing the essential properties of a physical, biological, or social system. Two very highly correlated variables may be assumed to be measuring the same underlying factor. Therefore, in the simplest case of factor analysis all variables may be divided into a few groups such that the correlation is very high for any two variables from the same group and very low for any two variables from different groups. The groups represent the factors. In more complex cases, some of the variables are composites of two or more factors.

6. MATHEMATICAL MODELS FOR STATISTICAL DATA

Interdependence among variables can be used to estimate any one of them as a dependent variable in terms of the others as independent variables. For qualitative variables the analysis of variance model is given by a general term plus a positive or negative increment associated with each main category of observations plus additional positive or negative increments due to interaction effects that may

arise from the simultaneous use of two or more classification systems to categorize the observations. For quantitative variables the regression model is given by a constant term plus a series of products, each a coefficient times an independent variable. By using antilogarithms the analysis of variance model may be transformed into the product of a general term times main effects times interaction effects. The regression model sum of products may be transformed into a product of powers in which the coefficients appear as exponents.

General linear hypothesis models incorporate both the analysis of variance for qualitative variables and regression analysis for quantitative variables. In addition to estimates for regression coefficients, general terms, category effects, and interactions, the general linear hypothesis also provides statistical tests to determine whether any one or a combination of these estimates differ significantly from zero. This permits the formulation of an optimum prediction function containing only those independent variables that significantly affect the value of the dependent variable.

All the definitions included in Section I for univariate statistics, probability densities, and measures of correlation could not be found in any single source. However, any one of them could be found in several other sources. For this reason specific references are not cited in the text. Instead a bibliography is given listing some of the more comprehensive sources from which more information may be obtained. Notations in the bibliography indicate the subject matter for which each reference is given.

SECTION II

STATISTICAL MEASURES FOR SINGLE VARIABLES

Measurements of a random variable require multiple observations because of unpredictable variations and errors associated with each measurement. Such a set of observations for any variable can be characterized by an overall average value as well as some measure of the dispersion or scatter in the data. In addition, whenever such a distribution of observed values is arranged in the bar graph form, characteristics related to the symmetry and shape of the distribution become evident. The concepts are quantified in the next few paragraphs.

Let x_i, i = 1...n denote a sample of n observations of the variable x. If the range of x is subdivided into a series of m adjacent intervals by the ordered sequence of values x_k, k = 0...m, m<<n then the function $y_k=f(x_k)$ can be defined to mean the number of observations that fall in the interval $x_{k-1}<x<x_k$. An example of such a function is plotted as a histogram in Figure 1a. For large numbers of observations and more refined subdivisions, such histograms approximate continuous mathematical functions illustrated in Figure 1b. Here the total area under the curve f(x) represents the total number of observations and the shaded area between the curve and any fixed interval on the x axis represents the number of observations expected to fall in that interval. If the area under the curve in Figure 1b is normalized to one by dividing f(x) by the total number of observations, the shaded area then represents the probability that a randomly selected observation x_i will fall in the underlying x interval. The probability density, p(x), the value of the

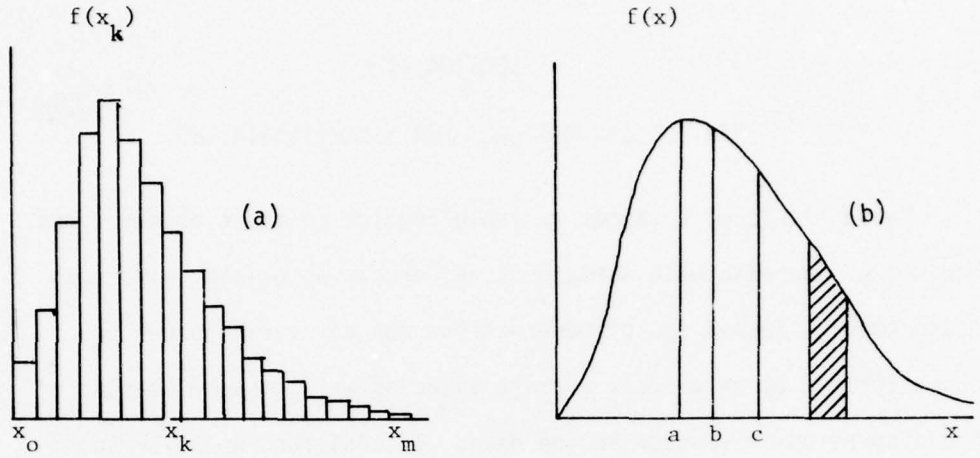


Figure 1. Multiple Observations of One Random Variable

ordinate in such a normalized curve, is defined as the limit of the ratio of the probability associated with a given interval to the length of the interval as the latter approaches zero. Probability density functions are therefore used to define infinite populations from which any given set of measurements constitutes only a small sample.

Statistics are numbers selected or computed from sample data to define various characteristics of the sample as a whole. Likewise, the mathematical expression for every probability density function (pdf) contains one or more constant parameters that define characteristics of the whole population. Sample statistics are therefore used to estimate unknown population parameters describing the same general characteristic. The characteristics of chief interest include measures of average value, measures of dispersion or scatter, and measures of the symmetry and shape of the histogram or probability density function. In the next three sections these characteristics are defined and in most cases computing formulas are given both for statistics in terms of summations of sample data and for parameters in terms of integrals of probability density functions.

1. MEASURES OF AVERAGE VALUES

Mode - The most frequently occurring value in a sample or population is called the mode. In Figure 1b, the mode is at point a, the value of x for which f(x) is a maximum.

Median - The midvalue in the sequence of observations ordered from lowest to highest is called the median. In Figure 1b, the median is at point b since a vertical line through b divides the area under the curve exactly in half.

Arithmetic Mean - The sum of all observations divided by the number of observations is called the arithmetic mean:

$$\bar{x} = \sum_{i=1}^{n} x_i / n \qquad \mu = \int_{-\infty}^{\infty} x p(x) dx \tag{1}$$

In Figure 1b, the mean is at point c, the x coordinate of the centroid (or balance point) of the area enclosed by the curve and the x axis.

Quadratic Mean - The square root of the mean of the squares of the observations is called the quadratic mean or the root mean square (rms):

$$\bar{x}_{rms} = \sqrt{\sum_{i=1}^{n} x_i^2 / n} \qquad \mu_{rms} = \sqrt{\int_{-\infty}^{\infty} x^2 p(x) dx} \tag{2}$$

The quadratic mean is always larger than the arithmetic mean and would be to the right of point c in Figure 1b.

Harmonic Mean - The reciprocal of the mean of the reciprocals of the observations is called the harmonic mean:

$$\bar{x}_{har} = \left[\sum_{i=1}^{n} x_i^{-1} / n \right]^{-1} \qquad \mu_{har} = \left[\int_{-\infty}^{\infty} x^{-1} p(x) dx \right]^{-1} \tag{3}$$

The harmonic mean is always less than the arithmetic mean and would be to the left of point c in Figure 1b.

Geometric Mean - The anti-logarithm of the mean of the logarithms of the observations is called the geometric mean:

$$\bar{x}_{geom} = \exp\left[\sum_{i=1}^{n} \ln x_i/n\right] \quad \mu_{geom} = \exp\int_0^\infty p(x) \ln x \, dx \qquad (4)$$

The Nth root of the product of all N observations is also the geometric mean since:

$$\bar{x}_{geom} = \exp\left[\sum_{i=1}^{n} \ln x_i/n\right] = \exp\left[\ln\prod_{i=1}^{n} x_i/n\right] = \sqrt[n]{\prod_{i=1}^{n} x_i} \qquad (4a)$$

The integral form does not exist in this case since an integral is the limit of a sum and there is no corresponding symbol for the limit of a product. The geometric mean is smaller than the arithmetic mean but larger than the harmonic mean for all positive observations.

Other Measures of Average Value - The quadratic, harmonic, and geometric means were defined by first taking functions of the observations - the square, reciprocal, and logarithm respectively; then computing the arithmetic mean; and finally taking the inverse functions - the square root, the reciprocal, and the anti-logarithm respectively. This same procedure can be employed by using other functions to generate definitions for other kinds of average values. The median values are unaffected by this process.

Selecting Appropriate Averages - The physical interpretation and application of the data are generally vital considerations in selecting the appropriate kind of average value to use. The character of the data itself will provide some guidelines. For data oscillating randomly about zero, the arithmetic mean is not useful since offsetting positive and negative fluctuations will always make it nearly zero irrespective of the magnitude of the oscillations - in this case the quadratic mean giving the root mean square value would be more useful. For data containing zeros and negative values the geometric mean is not suitable since even a single zero valued observation leads to a zero mean irrespective of other data, and an odd number of negative observations will lead to an imaginary mean if the total number of observations is even.

By selecting particular definitions of the mean the resulting value can be made larger or smaller almost at will. Consider equation 5 which defines the quadratic mean for m = 2 and the harmonic mean for $m = -1$

$$\bar{x} = \left[\sum_{i=1}^{n} x_i^m/n\right]^{1/m} \qquad \mu = \left[\int_{-\infty}^{\infty} x^m p(x)dx\right]^{1/m} \tag{5}$$

As m becomes increasingly positive the "mean" defined approaches the maximum value of x. As m becomes increasingly negative the "mean" defined approaches the minimum value of x. Consequently, by this or other less evident methods, a mean value definition can be selected to obtain a result almost anywhere in the range of the data being analyzed.

Selecting the correct kind of average to employ requires a correct understanding of the origin, interpretation, and application of the data involved. The following example is often cited: A vehicle travels 120 kilometers at 40 kilometers per hour and then another 120 kilometers at 60 kilometers per hour. What is the mean speed? The arithmetic mean (40+60)/2 = 50 is not correct. The vehicle requires three hours for the first segment and two hours for the second. Dividing the total 240 kilometers distance by the total five hour time gives an average of 48 kilometers per hour. This is the value of the harmonic mean: $[(40^{-1} + 60^{-1})/2]^{-1} = 48$. If the problem had been given with equal times rather than equal distances at the two speeds then the arithmetic mean would have given the correct value. For other kinds of problems a similar reasoning process must be employed. No general rule can be given for selecting the correct kind of average to use in every application.

2. MEASURES OF DISPERSION

Range - The difference between the largest and smallest values in a set of observations is called the range. In Figure 1a the sample range is $x_m - x_o$, in Figure 1b, the population range is infinite.

Mean Deviation - The mean absolute deviation of each observation from some average value for all observations is called the mean deviation. The median is the particular average value customarily used since the mean deviation is smaller about the median value than about any other number.

Standard Deviation - The square root of the mean squared deviation of each observation from the mean of all observations is called the standard deviation (std. dev.)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \sigma = \sqrt{\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx} \tag{6}$$

This root mean square deviation is smaller about the mean value than about any other number. Division by n-1 rather than n in Equation 6 compensates for this minimization bias that results from the use of the same sample to compute both mean and dispersion. If $\bar{x}$ were the mean of the population, the sum of squared deviations would be somewhat larger and the divisor would be n to obtain an unbiased estimate of the population standard deviation.

Variance - The mean square deviation about the mean is called the variance. It is the square of the standard deviation.

Coefficient of Variation - The ratio of the standard deviation to the arithmetic mean is called the coefficient of variation. Clearly it is a relative measure expressing the dispersion as a fraction of the mean value.

Other Dispersion Values - As with other mean values, definitions for other measures of dispersion can be generated by first taking functions of observations, then computing the selected measures of dispersion, and finally taking the inverse function of this result.

Selecting Appropriate Dispersions - As with selection of average values the physical interpretation and application of the data are generally vital considerations in selecting the appropriate measure of dispersion to use. Again as with selection of averages, the character of the data provides some guidelines - for example, avoiding relative measures of dispersion for data with negative or near zero mean values. With dispersions, however, one additional factor from sampling theory has a bearing. The extreme values, i.e., the maximum and minimum, from sets of measurements are subject to a very high degree of sampling variation. Consequently the range and other measures of dispersion using these extreme values are very unreliable as broad measures of dispersion for the data as a whole.

3. MEASURES OF SKEWNESS AND KURTOSIS

Both the mean and the variance of a random variable are related to a more general set of statistics called the moments of a probability density function. Moments are useful in specifying the shape of a pdf. The jth moment about point a is defined as follows for observed data and probability density functions, respectively:

$$\sum_{i=1}^{n} \frac{(x_i - a)^j}{n} \quad \text{and} \quad \int_{-\infty}^{\infty} (x - a)^j p(x)\,dx \qquad j = 1,2,\ldots \tag{7}$$

The first moment about zero is simply the arithmetic mean. The second moment about the mean is the variance. Third and fourth moments about the mean are used in computing coefficients of skewness and kurtosis which are associated with the symmetry and peakedness of a probability density function.

a. Standardized Data

In order to specify the higher moments in a more useful form, the effect of uniform changes in the observations on their mean value and standard deviation should be noted. Adding or subtracting a constant to each observation will add or subtract the same amount to the mean value but leave the standard deviation unchanged. On the other hand multiplying or dividing each observation by a constant value will multiply or divide both the mean value and the standard deviation by the same constant. In either kind of uniform change the relative magnitudes of the deviations from the mean and consequently the shape of the histogram or frequency distribution function remains unchanged.

A special case of uniform changes is of interest: first subtracting the mean value from each observation and then dividing the result by the standard deviation to form a new set of standardized data having a mean value of zero and a standard deviation of one. Standardized data are quite useful in studying characteristics of statistical data related to the shape of the probability density function and in measuring the correlation between two variables.

b. Moments of Standardized Data

The first four moments of the standardized value of x for both summations of sample data and integrals of probability density function are as follows:

First Moment

$$\frac{1}{n}\sum_{i=1}\left(\frac{x_i-\bar{x}}{s}\right) = \frac{1}{s}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{n} = 0$$

$$\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)p(x)dx = \frac{1}{\sigma}\int_{-\infty}^{\infty}(x-\mu)p(x)dx = 0$$

Second Moment

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^2 = \frac{1}{s^2}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n} = 1$$

$$\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^2 p(x)dx = \frac{1}{\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx = 1$$

Third Moment

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3 = \frac{1}{s^3}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{n} = \sqrt{b_1} = a_3 \qquad (8)$$

$$\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^3 p(x)dx = \frac{1}{\sigma^3}\int_{-\infty}^{\infty}(x-\mu)^3 p(x)dx = \sqrt{\beta_1} = a_3$$

Fourth Moment

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 = \frac{1}{s^4}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{n} = b_2 = a_4 \qquad (9)$$

$$\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)^4 p(x)dx = \frac{1}{\sigma^4}\int_{-\infty}^{\infty}(x-\mu)^4 p(x)dx = \beta_2 = a_4$$

Skewness - The standardized third moment, α_3, called the coefficient of skewness, is necessarily zero for symmetric probability density functions in which $p(x) = p(-x)$. (Sufficient conditions for symmetry require that all odd moments equal zero). For probability density functions skewed to the right as in Figure 1b, $\alpha_3 > 0$, for those skewed to the left $\alpha_3 < 0$.

Kurtosis - The standardized fourth moment, α_4, called the kurtosis, is associated with the varying degree of concentration about the mean that is possible for probability density functions having the same mean and standard deviation as shown in Figure 2.

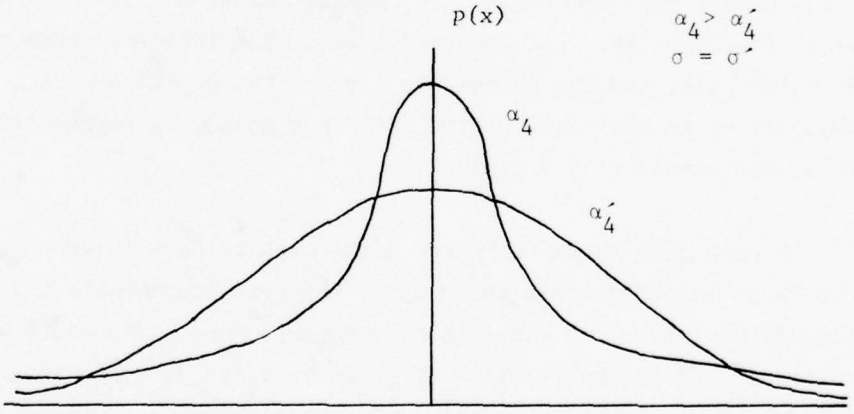


Figure 2. Kurtosis Variations in Probability Density Functions

Values of the skewness α_3 and the kurtosis α_4 for several probability densities of theoretical and practical interest are given in the next section.

SECTION III

PROBABILITY DENSITY FUNCTIONS FOR SINGLE VARIABLES

Square, sinusoidal, triangular, and random waves commonly used in instrumentation laboratories are each associated with probability density functions of theoretical interest as shown in Figure 3. The square wave amplitude lying alternately at the positive and negative extremes is represented by a discrete probability density function with delta functions at those extremes and zero elsewhere. The sine wave amplitude crossing the t axis at a steep angle and then flattening out at the extremes is represented by the U-shaped arc-sine distribution with a minimum at zero amplitude. The triangular wave, uniformly crossing all amplitudes within the wave range, is represented by the flat uniform distribution. The random wave amplitude, lying predominately near the t axis, is represented by the bell shaped normal distribution. For the random wave, the interval between a reference point and the nearest axis crossing, negative to the left and positive to the right of the reference point, is represented by the double exponential distribution.

In each case shown in Figure 3 the mean value μ is set equal to zero by so locating the t and y axes; the root mean square value (or standard deviation) is equal to σ by choosing the appropriate wave amplitude, and the coefficient of skewness α_3 is zero due to the symmetry with respect to positive and negative values. Therefore, illustrations of the kurtosis values α_4 indicated in Figure 3 are those associated with the corresponding pdf shapes. Each of those probability density functions may be given a nonzero mean, $\mu \neq 0$, by replacing x with $x - \mu$ in its mathematical expression. This corresponds to vertical shifts in the waveform and horizontal shifts in the density curve.

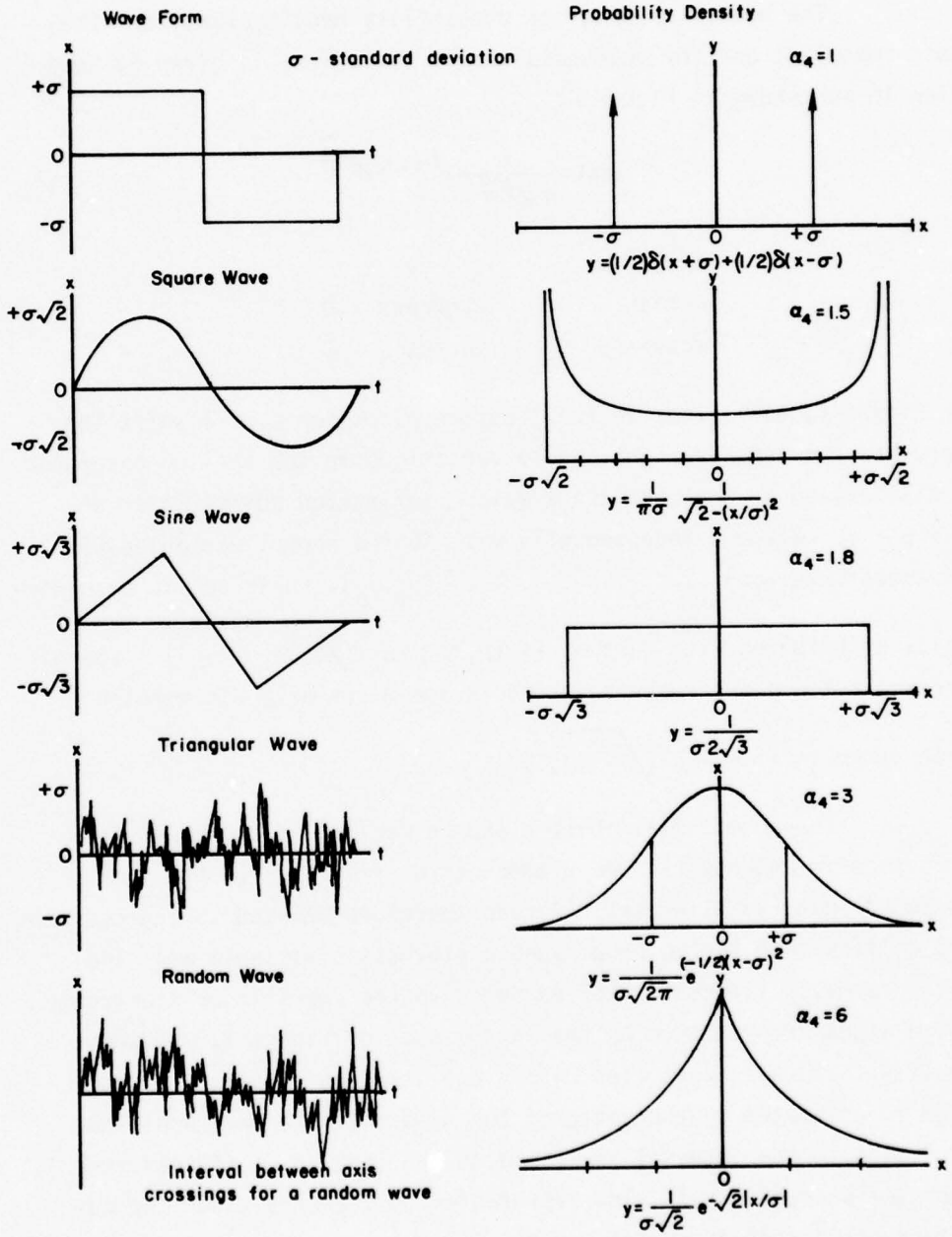


Figure 3. Wave Form Probability Densities

1. DISTRIBUTIONS UNBOUNDED ON BOTH SIDES

a. Normal Distribution

The normal or Gaussian probability density function, the most important one for unbounded random variables, is given by Equation 10 and shown in Figure 4:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-1/2)(\frac{x-\mu}{\sigma})^2} \tag{10}$$

mean = μ	st. dev = σ
median = μ	skewness = 0
mode = μ	kurtosis = 3

In Figure 4, variations in the location parameter μ will shift the curve to the left or right, while variations in the scaling parameter σ will expand or contract the x axis. Inflection points occur at $x = \mu \pm \sigma$. Given n independently distributed normal variables with parameters $(\mu_1,\sigma_1).....(\mu_n,\sigma_n)$, their sum is also normally distributed with parameters $(\mu_1+...+\mu_n, \sqrt{\sigma_1^2+...+\sigma_n^2})$. The difference between two such variables is again normally distributed with parameters $(\mu_1-\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$.

The normal distribution can be derived as the model for processes in which total measurement error results from the summation of many small errors. It can therefore be used to express the probability that the value of such a stochastic variable will lie within a given interval. For example, in the acoustic or electronic noise signal represented by the random wave of Figure 3, the probability that the signal lies within one standard deviation of the mean is estimated by the ratio of the time that the wave amplitude $x(t)$ lies in the interval $-\sigma \leq x(t) \leq \sigma$ to the total time of measurement. The same probability is also represented by the area under the adjacent normal pdf curve between the lines $x = -\sigma$ and $x = +\sigma$.

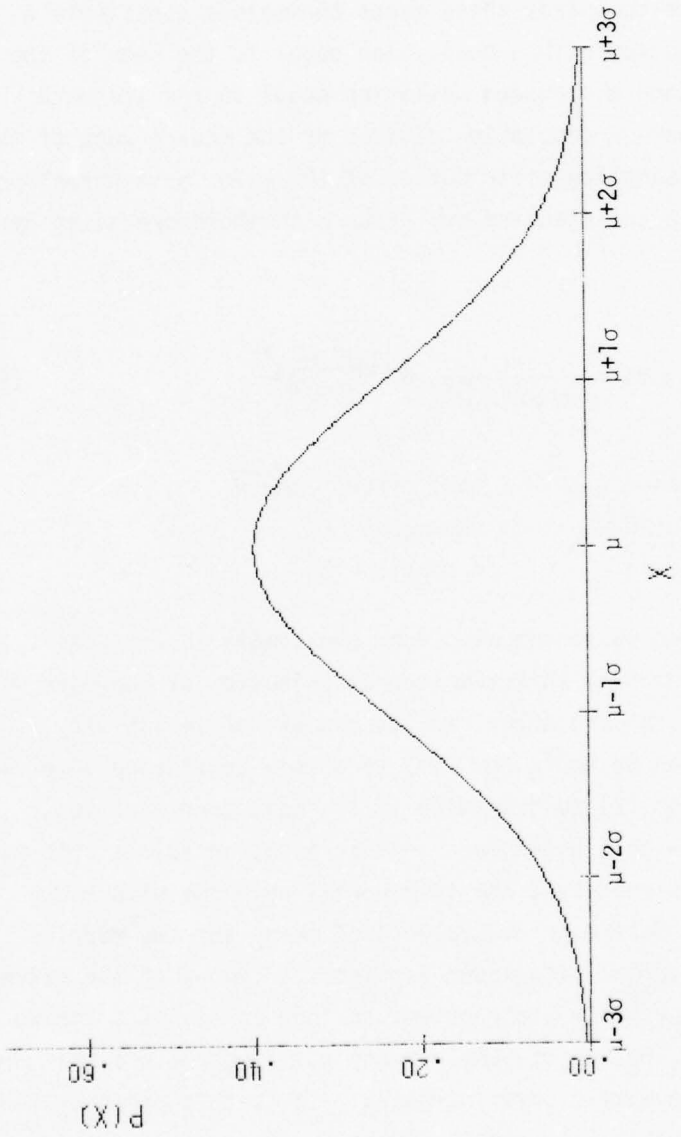


Figure 4. Normal Probability Density

b. Sampling Distribution of the Mean

The normal pdf is not only the basic distribution law for many stochastic processes but also plays a vital role in statistics. If samples of size n are repeatedly drawn from a normal population and mean values are computed, these means themselves constitute a new normal distribution with a mean value equal to the mean of the parent population and a standard deviation equal to the standard deviation of the parent population divided by the square root of the sample size. The sampling distribution of the mean for a normal population with mean μ and standard deviation σ is therefore given by Equation 10a:

$$p(x) = \frac{1}{(\sigma/\sqrt{n})\sqrt{2\pi}} e^{(-1/2)\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2} \qquad (10a)$$

mean = μ	std. error = $\sigma/\sqrt{n}$
median = μ	skewness = 0
mode = μ	kurtosis = 3

Consequently, a mean value computed from one sample is, in reality, a single observation from this sampling distribution of the mean with $\sigma/\sqrt{n}$, called the standard error, as its measure of variability. Accordingly, 10a can be employed: (1) to obtain confidence intervals for sample means, or (2) to determine if a sample mean represents a normal universe with a known mean and variance, or (3) to determine if two sample means represent the same normal universe with known variance or (4) to determine equivalence of means for two samples from different universes with known variances. Even when the parent population is not normally distributed, so long as it has a finite variance, the distribution of sample means still approaches the normal distribution as the sample size increases. It is this property that gives the normal probability density function its position of importance in statistics.

c. Student t Distribution

Both the mean and the standard deviation of a normal population must be known to obtain the distribution of sample means using Equation 10a. If the mean is given but the standard deviation must be estimated from the sample, a simple substitution of s for σ is not sufficient. This is because $(\bar{x}-\mu)/(\sigma/\sqrt{n})$ is normally distributed with zero mean and unit variance but the exact distribution of $(\bar{x}-\mu)/(s/\sqrt{n})$ depends on the sample size n. The Student t distribution, given by Equation 11 and shown in Figure 5 is the sampling distribution of

$$t=(x-u)/(s/\sqrt{n}):$$

$$p(t)=\frac{\Gamma(\frac{1}{2}+\frac{\nu}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})\sqrt{\nu}}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \qquad \nu=n-1 \tag{11}$$

mean = 0	std. dev. = $\sqrt{\nu/(\nu-2)}$	$\nu > 2$
median = 0	skewness = 0	$\nu > 3$
mode = 0	kurtosis = $3 + 6/(\nu-4)$	$\nu > 4$

Inflection points occur at $\pm\sqrt{\nu/(\nu+2)}$

The t distribution approaches the unit normal as $\nu\to\infty$. The definition of t and Equation 11 can be used (1) to obtain confidence intervals for population means, or (2) to determine if a sample mean represents a normal universe with known mean but unknown variance, or (3) to determine if two sample means represent the same normal universe with an unknown variance, or (4) to determine equivalence of means for two samples from different normal universes with unknown variances. Application of the Student t distribution for such hypothesis tests will be treated in Section IV, Significance Tests for Pairs of Means and Variances.

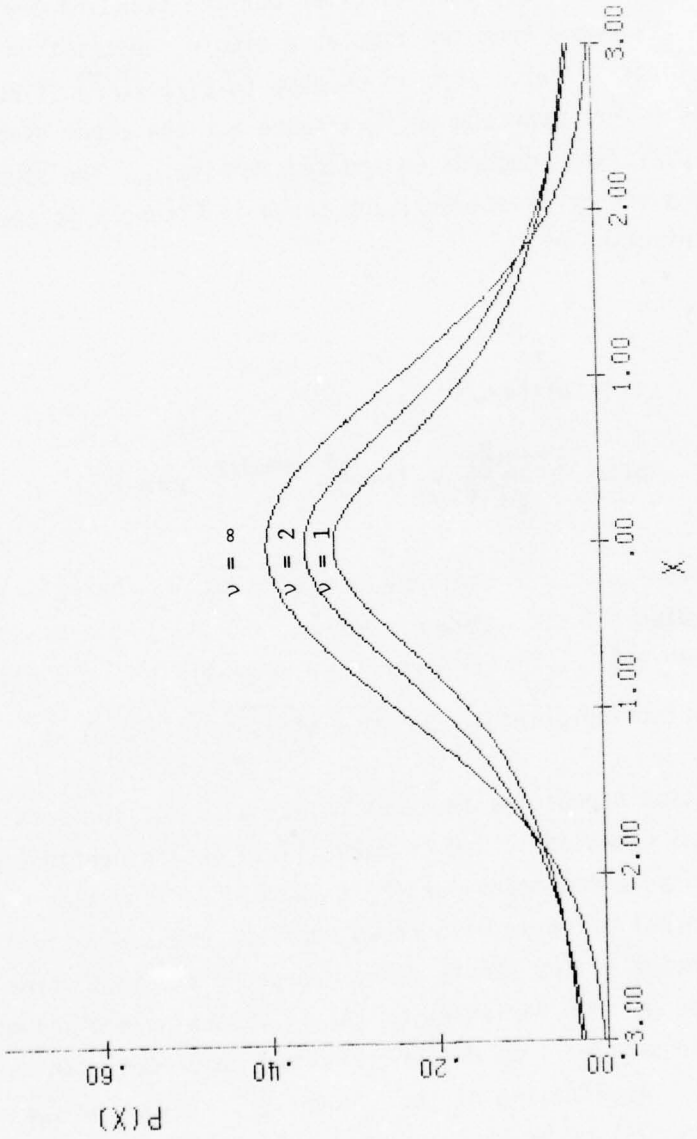


Figure 5. Student t Probability Density

d. Cauchy Distribution

A distribution of theoretical interest is obtained from the Student t distribution by taking $\nu = 1$ and setting $t = (x-\mu)/\sigma$ with μ and σ serving respectively as location and scaling parameters (but not as mean and standard deviation since all moments are infinite). The Cauchy distribution, given by Equation 12 and shown in Figure 5 for $\nu = 1$ has no finite mean or standard deviation:

$$p(x) = \frac{1}{\sigma\pi} \quad \frac{1}{1+\left[(x-\mu)/\sigma\right]^2}$$

$$\int_{-\infty}^{x} p(x)\,dx = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-\mu}{\sigma}\right) \tag{12}$$

median = μ mode = μ

Inflection points occur at $\mu \pm \sigma/\sqrt{3}$

The ratio of observations from two independent normal distributions having zero means and standard deviations σ_1 and σ_2 is Cauchy distributed with $\mu=0$ and $\sigma=\sigma_1/\sigma_2$. Since this ratio can be inverted, the reciprocal of a Cauchy variate is also Cauchy distributed. The sum of independent Cauchy variates is also Cauchy distributed. Consequently, the arithmetic mean has the same Cauchy distribution as the individual observations. Previously, it was noted that the distribution of the arithmetic mean approaches the normal as sample size increases, no matter how the parent population is distributed so long as it has finite variance. The Cauchy distribution does not meet the last condition. Because of its infinite variance, the mean of a Cauchy pdf is no more informative than a single observation.

2. DISTRIBUTIONS BOUNDED ON ONE SIDE

a. Gamma Distribution

The gamma probability density function, an important one for random variables bounded on one side, is given by Equation 13 and shown in Figure 6:

$$p(x) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x-\gamma}{\beta}\right)} \qquad \alpha>0, \beta>0, x>\gamma \tag{13}$$

mean value = $\gamma + \alpha\beta$ skewness = $2/\sqrt{\alpha}$

std. dev. = $\beta\sqrt{\alpha}$ kurtosis = $3 + 6/\alpha$

Gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$; $\Gamma(1) = 1, \Gamma(\frac{1}{2}) = \sqrt{\pi}$

recurrence formula $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

for α an integer $\Gamma(\alpha) = (\alpha-1)\ (\alpha-2)(\alpha-3).....(1) = (\alpha-1)!$

In Figure 6, variations in the location parameter γ will shift the curves to the left or right, while variations in the scaling parameter β will expand or contract the x axis. For $\alpha \geq 1$, the curves have a mode at $x = \gamma + \beta(\alpha-1)$ = mean - β. Inflection points equidistant from the mode occur at $x = \gamma + \beta[(\alpha-1) \pm \sqrt{(\alpha-1)}]$ = (mean - β) $\pm\ \beta\sqrt{\alpha-1}$. As the shape parameter α increases, the skewness and kurtosis approach 0 and 3 respectively and the curve shape becomes more like the normal. If $x_1, \ldots, x_n$ are independent random variables having gamma distributions with common values of β and γ, then their sum $(x_1+.....+x_n)$ also has a gamma distribution with the same values of β and γ and with $\alpha = \alpha_1+...+\alpha_n$.

The gamma distribution is the appropriate model for the time required for a total of α independent events to occur if the mean time per event β remains constant. An example would be α axis crossings of the random wave in Figure 3. The time for one event or equivalently the time between events is therefore a special case called the exponential distribution.

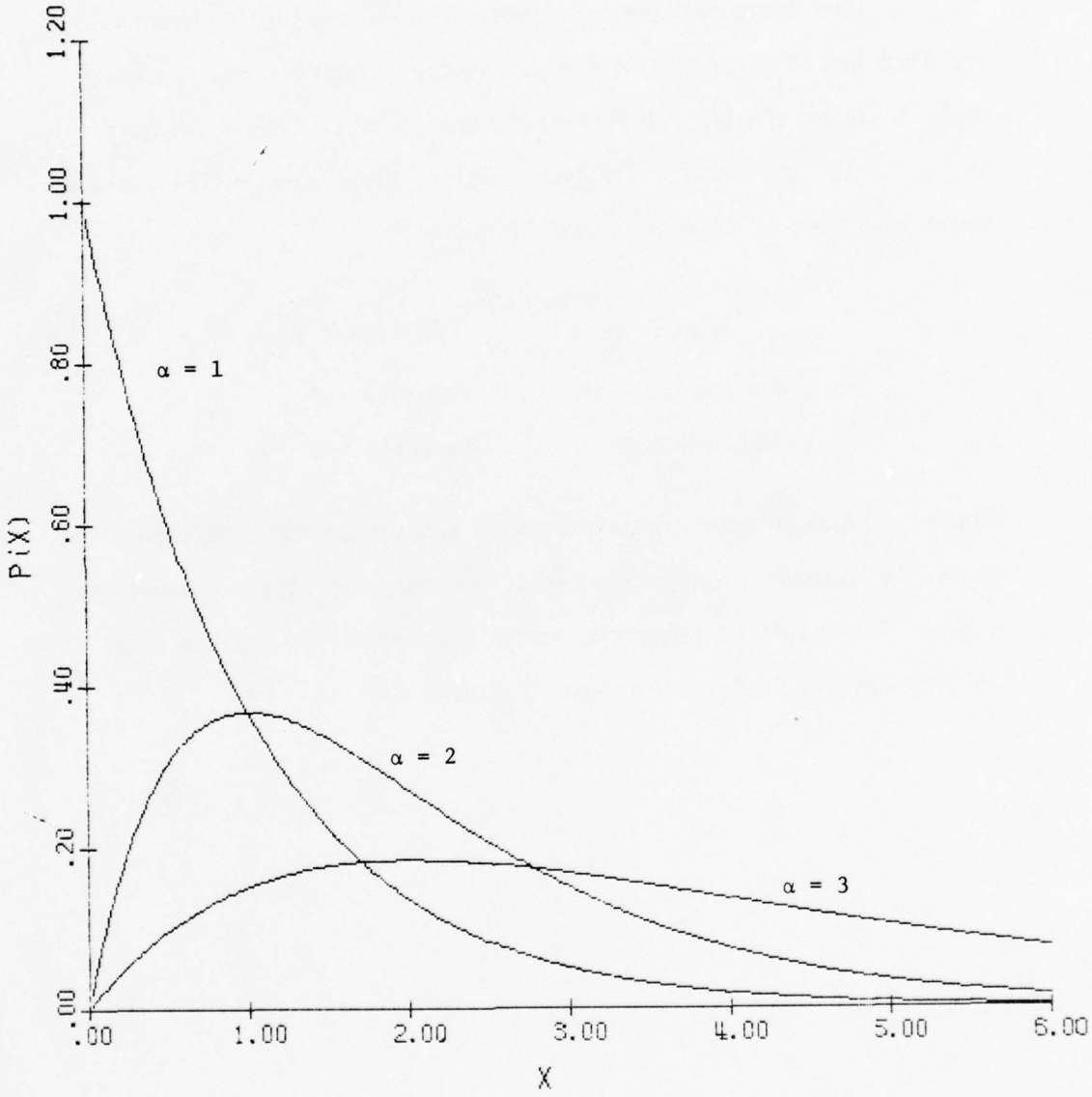


Figure 6. Gamma Probability Density Function

b. Exponential Distribution

The times between independent randomly occurring events are distributed according to the exponential distribution. An example would be the times between axis crossings of the random wave in Figure 3. The exponential distribution, given by Equation 13a and shown in Figure 6, is a gamma pdf with $\alpha = 1$:

$$p(x) = \frac{1}{\beta} e^{-(x-\gamma)/\beta} \qquad \beta > 0, x > \gamma \tag{13a}$$

mean value = $\gamma+\beta$ skewness = 2

std. dev = β kurtosis = 9

Since the times between events preceding and succeeding a reference point are measured in opposite directions they may be given opposite signs. The resulting symmetric double exponential or Laplace probability density function is shown in Figure 3.

c. Chi-Square Distribution

If random variable x is normally distributed with mean μ and standard deviation σ, then chi-square (χ^2) with n degrees of freedom is defined as the sum of squares of n standardized normal observations:

$$\chi^2 = \sum_{i=1}^{n} \left[(x_i - \mu)/\sigma \right]^2$$

The chi-square distribution, given by Equation 13b, is a gamma pdf with $\alpha=n/2$, $\beta=2$, and $\gamma=0$:

$$p(\chi^2) = \frac{1}{2\Gamma\left(\frac{n}{2}\right)} \left(\frac{\chi^2}{2}\right)^{\frac{n}{2}-1} e^{-\chi^2/2} \qquad \chi^2 \geq 0 \qquad (13b)$$

mean value = n skewness = $2/\sqrt{n/2}$

std. dev. = $\sqrt{2n}$ kurtosis = 3 + 12/n

For the case n=1, this is the distribution of the squares of standardized normal data which has mean=1, st. dev. = $\sqrt{2}$, skewness = $2\sqrt{2}$, and kurtosis = 15. The chi-square distribution plays a vital role in statistics because simple transformations of it yield important sampling distribution as noted in the next four paragraphs.

d. Sampling Distribution of the Variance and Variance Ratio

Multiplying chi-square by σ^2/n gives the variance for a sample of size n from a normal universe with mean μ and standard deviation σ:

$$\frac{\sigma^2}{n}\chi^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 = s^2$$

The sampling distribution of the variance s^2 given by Equation 13c is a gamma pdf with $\alpha = n/2$, $\beta = 2\sigma^2/n$, and $\gamma = 0$:

$$p(s^2) = \frac{1}{(2\sigma^2/n)\,\Gamma\left(\frac{n}{2}\right)}\left(\frac{s^2}{2\sigma^2/n}\right)^{\frac{n}{2}-1} e^{-\frac{s^2}{2\sigma^2/n}} \qquad s^2 \geq 0 \quad (13c)$$

mean value = σ^2 skewness = $2/\sqrt{n/2}$

std. dev. = $\sigma^2/\sqrt{n/2}$ kurtosis = $3 + 12/n$

Variances computed from independent samples of the same normal population constitute a gamma pdf having a mean equal to the population variance and a standard deviation equal to the population variance divided by the square root of half the sample size.

The sample to population variance ratio s^2/σ^2, given by χ^2/n, is frequently more convenient to use:

$$\frac{\chi^2}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2 = \frac{s^2}{\sigma^2}$$

The sampling distribution of the variance ratio s^2/σ^2, given by Equation 13d, is a gamma pdf with $\alpha=n/2$, $\beta=2/n$, and $\gamma=0$:

$$p\left(\frac{s^2}{\sigma^2}\right) = \frac{1}{(2/n)\Gamma(n/2)}\left(\frac{s^2/\sigma^2}{2/n}\right)^{\frac{n}{2}-1} e^{-\frac{s^2/\sigma^2}{2/n}} \qquad s^2/\sigma^2 \geq 0 \qquad (13d)$$

mean value = 1 skewness = $2/\sqrt{n/2}$

std. dev. = $\sqrt{2/n}$ kurtosis = 3 + 12/n

This chi-square over degrees of freedom probability density function is commonly employed in statistical theory to obtain confidence intervals for variances, or to determine if a sample variance is consistent with some fixed population value suggested by theory.

e. F Distribution

The distribution of the ratio of two variates having independent chi-square over degrees of freedom probability densities is also the distribution of the ratio F of sample variances from two normal universes:

$$\frac{1}{n_j-1}\chi_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(\frac{x_{ji}-\bar{x}_j}{\sigma_j}\right)^2 = \frac{1}{\sigma_j^2}\sum_{i=1}^{n_j}\left(x_{ji}-\bar{x}_j\right)^2/\left(n_j-1\right) = s_j^2/\sigma_j^2$$

$$\frac{\chi_1^2/\left(n_1-1\right)}{\chi_2^2/\left(n_2-1\right)} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2}\frac{s_1^2}{s_2^2} = \frac{\sigma_2^2}{\sigma_1^2}F$$

The F distribution given by Equation 14 and shown in Figure 7 is the sampling distribution of the ratio of two sample variances, $F = s_1^2/s_2^2$ having ν_1 and ν_2 degrees of freedom

$$p(F) = \nu_1^{\nu_1/2}\,\nu_2^{\nu_2/2}\,\frac{\Gamma\left(\frac{\nu_1}{2}+\frac{\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\,\frac{F^{(\nu_1/2)-1}}{\left(\nu_2+\nu_1 F\right)^{(\nu_1+\nu_2)/2}} \qquad (14)$$

$$\text{mean value} = \nu_2/(\nu_2-2) \qquad \text{skewness} = \frac{(2\nu_1+\nu_2-2)}{(\nu_2-6)}\sqrt{\frac{8(\nu_2-4)}{(\nu_1+\nu_2-2)\nu_1}}$$

$$\text{std dev} = \frac{\nu_2\sqrt{2(\nu_1+\nu_2-2)}}{(\nu_2-2)\sqrt{\nu_1(\nu_2-4)}} \qquad \text{kurtosis} = 3+\frac{12\left[(\nu_2-2)^2(\nu_2-4)+\nu_1(\nu_1+\nu_2-2)(5\nu_2-22)\right]}{\nu_1(\nu_2-6)(\nu_2-8)(\nu_1+\nu_2-2)}$$

Mean value, std. dev., skewness and kurtosis formulas hold only for $\nu_2 > 2$, $\nu_2 > 4$, $\nu_2 > 6$ and $\nu_2 > 8$ respectively.

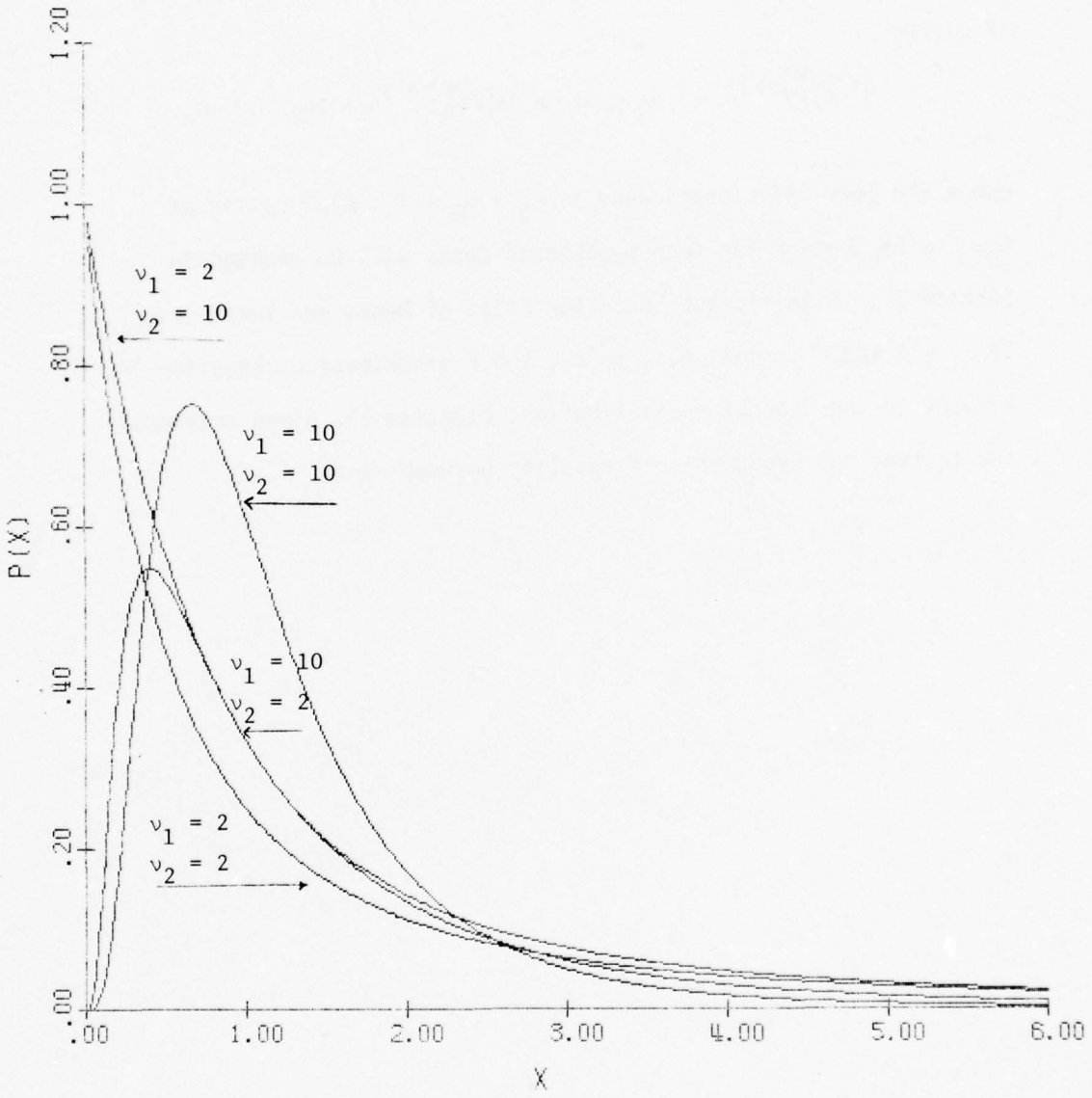


Figure 7. F Probability Density Function

The F distribution is used in testing the hypothesis of equality between variances σ_1^2 and σ_2^2. Specifically $\sigma_1 = \sigma_2$ would be rejected if either

$$\left(s_1^2/s_2^2\right) < F_{n_1-1,\,n_2-1,\,\alpha_1} \text{ or } \left(s_1^2/s_2^2\right) > F_{n_1-1,\,n_2-1,\,1-\alpha_2}$$

where the level of significance is $\alpha_1 + \alpha_2 < 1$. Application of the F distribution for such hypothesis tests will be treated in Section IV, Significance Tests for Pairs of Means and Variances. If $\nu_1 = 1$ and F is set equal to t^2, the F distribution, Equation 14, reduces to the Student t distribution, Equation 11, given previously for testing the hypothesis of equality between means.

f. Rayleigh and Maxwell Distributions

Multiplying chi-square by σ^2 gives the error sum of squares for a sample of size n from a normal population with mean μ and standard deviation σ:

$$\sigma^2\chi^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} (x_i-\mu)^2$$

The Rayleigh distribution, given by Equation 15 and shown in Figure 8, is the probability density of $\sqrt{\sigma^2\chi^2} = x$ for n = 2

$$p(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} \quad x \geq 0 \tag{15}$$

mean value = $\sigma\sqrt{\pi/2} = 1.25\sigma$ skewness = $\sqrt{\pi/2}\,(\pi-3)/(2-\pi/2)^{3/2} = 0.2709$

std. dev. = $\sigma\sqrt{2-\pi/2} = 0.66\sigma$ kurtosis = $(8-3\pi^2/4)/(2-\pi/2)^2 = 3.2451$

The Maxwell distribution, given by Equation 16 and shown in Figure 9, is the probability density of $\sqrt{\sigma^2\chi^2} = x$ for n = 3

$$p(x) = \sqrt{\frac{2}{\pi}} \frac{x^2}{\sigma^3} e^{-x^2/2\sigma^2} \quad x \geq 0 \tag{16}$$

mean value = $2\sigma\sqrt{2/\pi} = 1.60\sigma$ skewness = $\sqrt{2/\pi}\,(32/\pi-10)/(3-8/\pi)^{3/2} = 0.4857$

std. dev. = $\sigma\sqrt{3-8/\pi} = 0.67\sigma$ kurtosis = $(15+16/\pi-192/\pi^2)/(3-8/\pi)^2 = 3.1082$

If the errors in the coordinates of a rectangular system are independent and normally distributed with the same variance, then the distribution of radial error is Rayleigh for a plane and Maxwell for a volume. Similarly, if the rectangular components of a particle velocity are independent and normally distributed with the same variance, then the distribution of particle speed is Rayleigh for motion in a plane and Maxwell for motion in a volume.

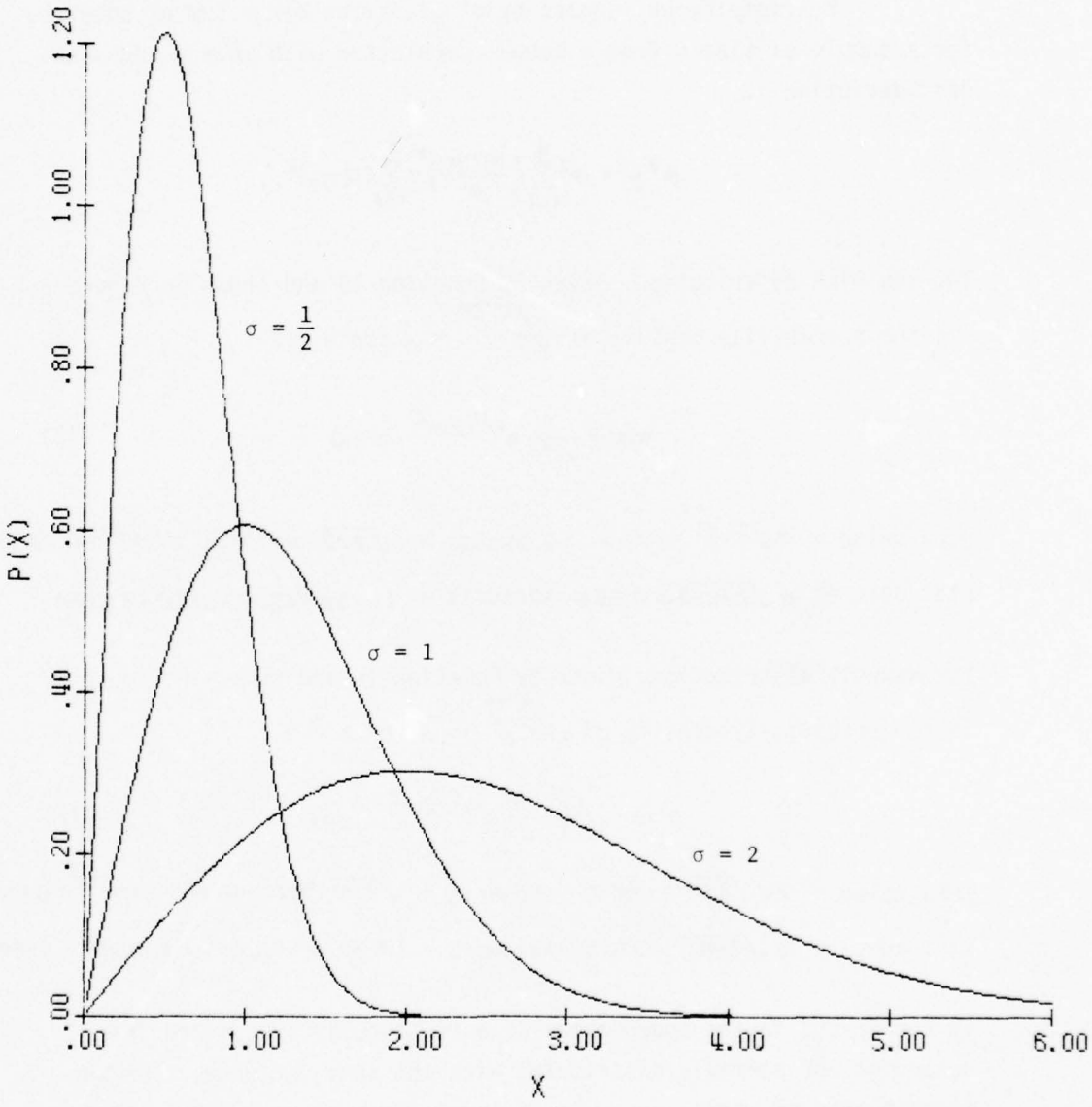


Figure 8. Rayleigh Probability Density Function

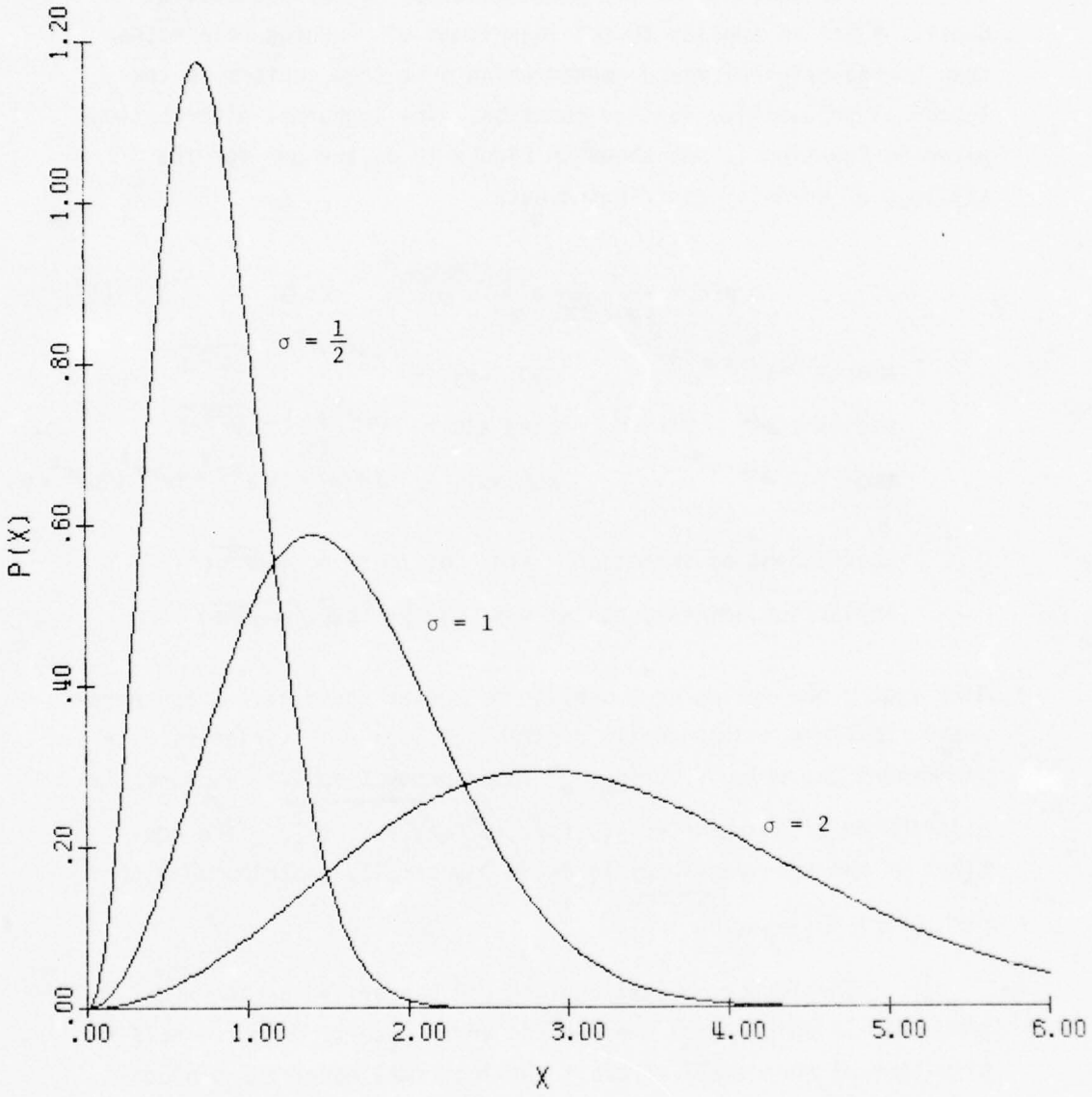


Figure 9. Maxwell Probability Density Function

g. Lognormal Distribution

For some stochastic processes, the normal probability density function applies to the logarithms of measured data rather than to the original measurements which must then conform to the lognormal probability density function. The lognormal distribution given by Equation 17 and shown in Figure 10 is the pdf for the antilogs of normally distributed data.

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2} \qquad x>0 \tag{17}$$

mean = $e^{\mu+\sigma^2/2}$ std. dev. = $e^{\mu+\sigma^2/2}\sqrt{e^{\sigma^2}-1}$

median = e^{μ} skewness = $(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$

mode = $e^{\mu-\sigma^2}$ kurtosis = $3+(e^{\sigma^2}-1)(e^{3\sigma^2}+3e^{2\sigma^2}+6e^{\sigma^2}+6)$

Coefficient of variation = std. dev./mean = $\sqrt{e^{\sigma^2}-1}$

Inflection points occur at exp $(\mu-\frac{3}{2}\sigma^2\pm\sigma\sqrt{1+\sigma^2/4})$

Note that μ now serves as a scaling parameter and σ as a shape parameter. Given n independently distributed lognormal variables with parameters $(\mu_1,\sigma_1).....(\mu_n,\sigma_n)$ their product is also lognormally distributed with parameters $(\mu_1+....+\mu_n,\sqrt{\sigma_1^2+....+\sigma_n^2})$. The quotient of two such variables is again lognormally distributed with parameters $(\mu_1-\mu_2,\sqrt{\sigma_1^2+\sigma_2^2})$.

The lognormal distribution can be derived as the model for processes in which total measurement error results from the multiplication of many small errors. The lognormal model for products is thus analogous to the normal model for sums. In particular, if samples of size n are repeatedly drawn from a lognormal population and geometric means are computed, these geometric means themselves constitute a new lognormal distribution with parameters $(\mu, \sigma/\sqrt{n})$.

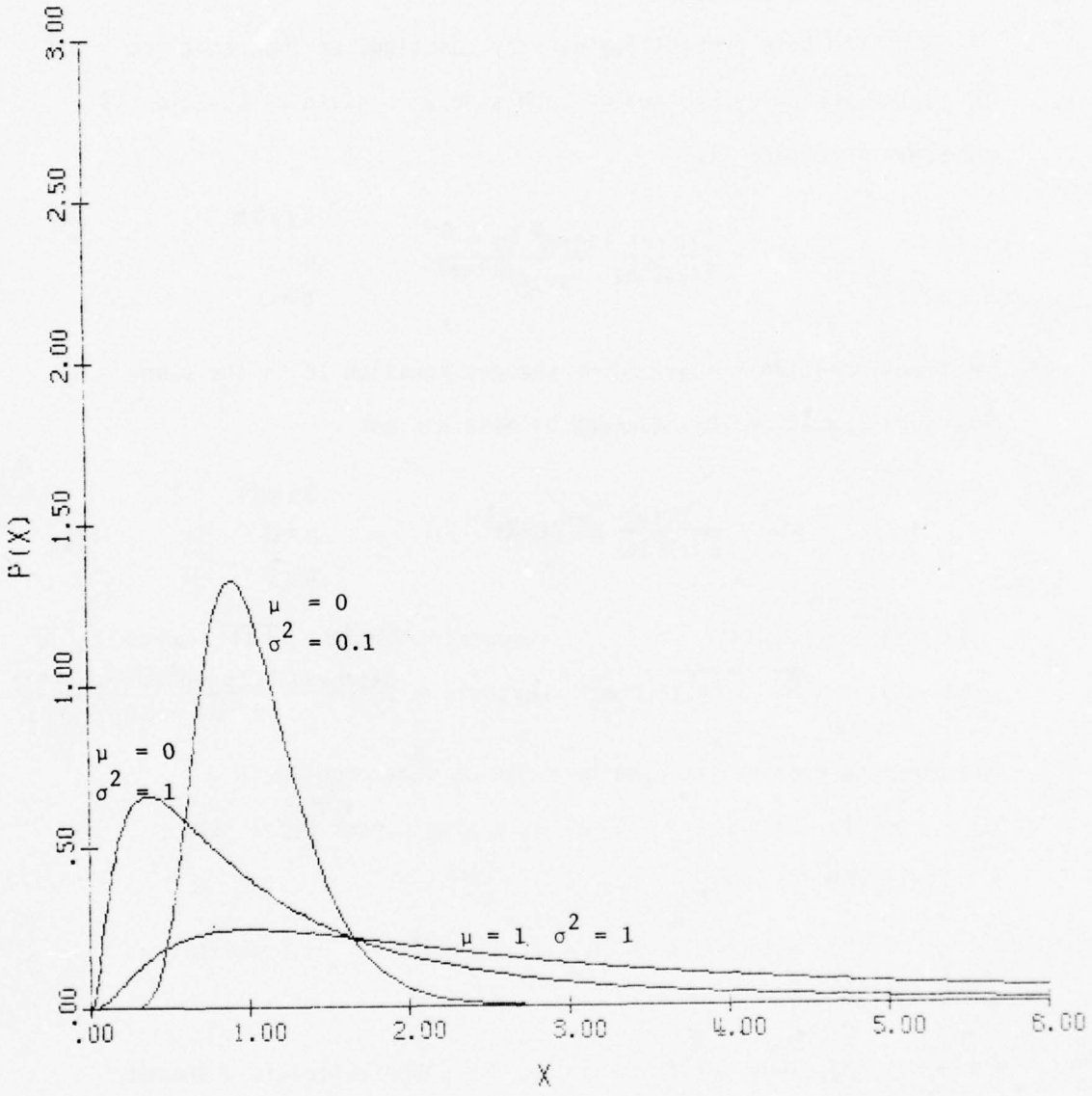


Figure 10. Lognormal Probability Density Function

3. DISTRIBUTIONS BOUNDED ON BOTH SIDES

a. Beta Distribution

The beta probability density function, an important one for random variables bounded on both sides, is given by Equation 18 and shown in Figure 11:

$$p(y) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{(y-a)^{p-1}(b-y)^{q-1}}{(b-a)^{p+q-1}} \qquad \begin{matrix} a \le y \le b \\ p > 0 \\ q > 0 \end{matrix} \qquad (18)$$

The transformation $x = (y-a)/(b-a)$ changes Equation 18 to the standard form (Equation 18a) bounded by zero and one

$$p(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1} \qquad \begin{matrix} 0 \le x \le 1 \\ p > 0 \\ q > 0 \end{matrix} \qquad (18a)$$

mean value $= p/(p+q)$ skewness $= \dfrac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$

std. dev. $= \sqrt{pq/(p+q+1)}/(p+q)$ kurtosis $= \dfrac{3(p+q+1)[2(p+q)^2+pq(p+q-6)]}{pq(p+q+2)(p+q+3)}$

For the beta probability density function (Equation 18a):

(1) When $p > 1$ and $q > 1$, a single peak occurs at $x = (p-1)/(p+q-2)$.

(2) When $p < 1$ and $q < 1$, a single valley occurs at $x = (p-1)/p+q-2)$.

(3) When $p \geq 1$ and $q < 1$, the distribution is J shaped.

(4) When $p < 1$ and $q \geq 1$, the distribution is reverse J shaped.

(5) When $p = q$, the distribution is symmetrical.

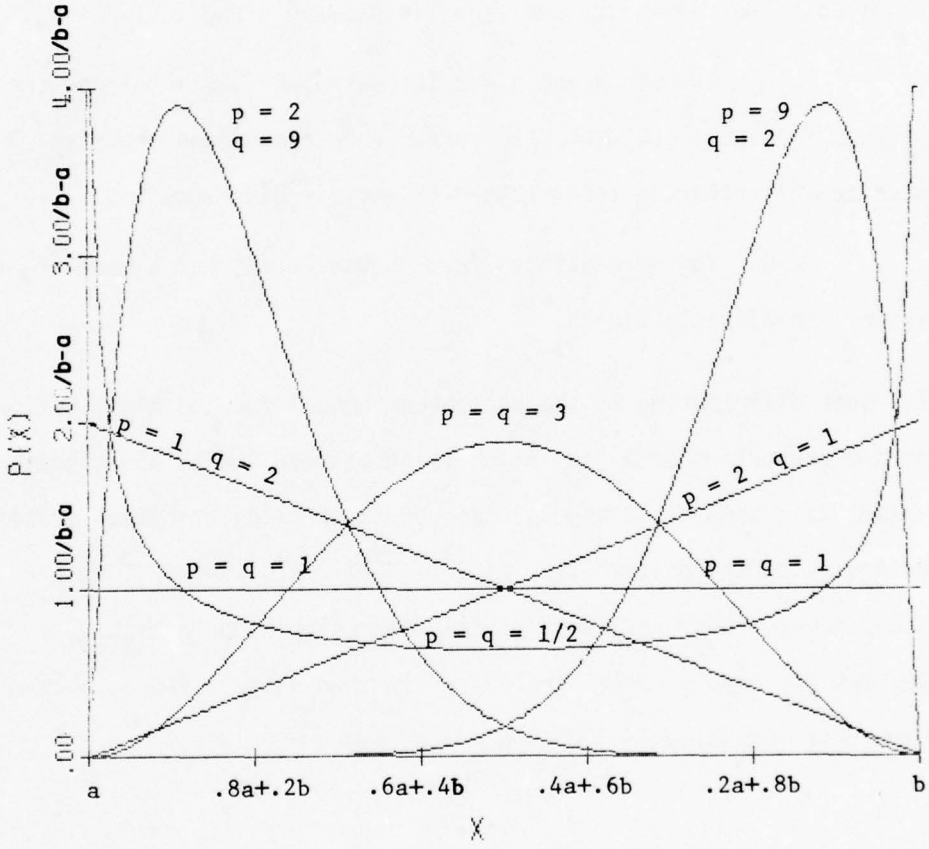


Figure 11. Beta Probability Density Function

(6) For all positive values of p and q, there are points of inflection at

$$\frac{p-1}{p+q-2} \pm \frac{1}{p+q-2}\sqrt{\frac{(p-1)(q-1)}{p+q-3}}$$

provided these values are real and lie between 0 and 1.

(7) If both p and q are increased while maintaining the ratio $\mu = p/(p+q)$ constant, the variance decreases and the standardized distribution tends toward the normal distribution.

(8) The more general form (Equation 18) has a mode at $x = a + (b-a)(p+q)/(p+q-2)$.

The beta distribution is the appropriate model for the distribution of the proportion of a population lying between lowest and highest values in a sample. A special case of the beta distribution arises naturally as the distribution of $V^2 = X_1^2/(X_1^2+X_2^2)$ where X_1^2, X_2^2 are independent random variables distributed as chi-square with ν_1, ν_2 degrees of freedom respectively. V^2 is then distributed as a standard beta (Equation 18a) with $p = \nu_1/2$ and $q = \nu_2/2$.

b. Uniform Distribution

The triangular wave with a random phase of starting point has a uniform probability density function as shown in Figure 3. The uniform distribution, given by Equation 18b and shown as the horizontal line in Figure 11, is a beta pdf with p=1 and q=1:

$$p(x) = 1/(b-a) \qquad a \leq x \leq b \tag{18b}$$

mean value = (a + b)/2 skewness = 0

std. dev. = $(b - a)/2\sqrt{3}$ kurtosis = 1.8

Solving for a and b in terms of the mean μ and standard deviation σ and substituting into p(x) gives the uniform pdf in the form shown in Figure 3.

c. Arc-Sine Distribution

The sine wave with a random phase or starting point has a U-shaped arc-sine probability density function as shown in Figure 3. The arc-sine distribution, given by Equation 18c and shown as the U-shaped curve in Figure 11 is a beta pdf with p=1/2, q=1/2, and $-a \leq x \leq a$

$$p(x) = 1/\pi\sqrt{a^2 - x^2} \qquad -a \leq x \leq a \tag{18c}$$

mean value = 0 skewness = 0

std. dev. = $a/\sqrt{2}$ kurtosis = 1.5

Setting $a = \sigma\sqrt{2}$ in p(x) gives the arc-sine pdf in the form shown in Figure 3. The name arc-sine comes from the cumulative form of this pdf given by Equation 18d

$$\int p(x)\,dx = \int_{-a}^{x} \left(\pi\sqrt{a^2 - x^2}\right)^{-1} dx = \pi^{-1}\arcsin\frac{x}{a} + \frac{1}{2} \tag{18d}$$

SECTION IV

SIGNIFICANCE TESTS FOR PAIRS OF MEANS AND VARIANCES

Statistical tests of significance are used to determine whether the same statistical quantities computed from two different samples differ by more than would be expected from sampling variations alone. If they do, the conditions under which the two samples were obtained have produced significant effects which must be accounted for in subsequent analysis. The two most common statistical tests are the F test and the t test used, respectively, to test for equality of variances and equality of means computed from two different samples.

1. TEST FOR EQUALITY OF VARIANCES

The F statistic is given by the formula

$$F = \frac{s_1^2}{s_2^2} \quad \text{where} \quad \begin{array}{l} s_1 = \text{larger of two standard deviation values} \\ s_2 = \text{smaller of two standard deviation values} \end{array} \qquad (19)$$

Significant differences in variance exist if this computed F exceeds the tabulated F that will be found in the row and column headed respectively by one less than the denominator sample size and one less than the numerator sample size. Such standard F tables will be found in almost any statistics text for several choices of the desired level of significance, i.e., the probability of rejecting the equality of variance hypothesis when it is, in fact, true.

2. TEST FOR EQUALITY OF MEANS

The t test for equality of means for the two samples is of two forms depending on whether the sample variances are equal or not.

If the sample variances are equal ($\sigma_1^2=\sigma_2^2$):

$$t = \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad \text{where} \quad s_p = \frac{\sum x_{1i}^2-\frac{(\sum x_{1i})^2}{n_1}+\sum x_{2i}^2-\frac{(\sum x_{2i})^2}{n_2}}{n_1+n_2-2} \tag{20}$$

If the sample variances are not equal ($\sigma_1^2\neq\sigma_2^2$):

$$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{s_1^2/n_1+s_2^2/n_2}} \quad \text{with} \quad df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1+1}+\frac{(s_2^2/n_2)^2}{n_2+1}}-2 \tag{21}$$

where x_{1i}, x_{2i} = sample values

$\bar{x}_1$, $\bar{x}_2$ = sample means

n_1 , n_2 = sample sizes

s_1, s_2 = sample standard deviations

df = degrees of freedom

For the equal variance case, significant differences in means exist if the computed t exceeds in absolute value the tabulated t for $n_1 + n_2 - 2$ degrees of freedom. For the unequal variance case, significant differences in means exist if the computed t exceeds in absolute value the tabulated t for the degrees of freedom given by the expression for df. Standard t tables will be found in almost any statistics text for several choices of the desired level of significance (the probability of rejecting the equality of means if true).

3. CHOICE OF LEVEL OF SIGNIFICANCE

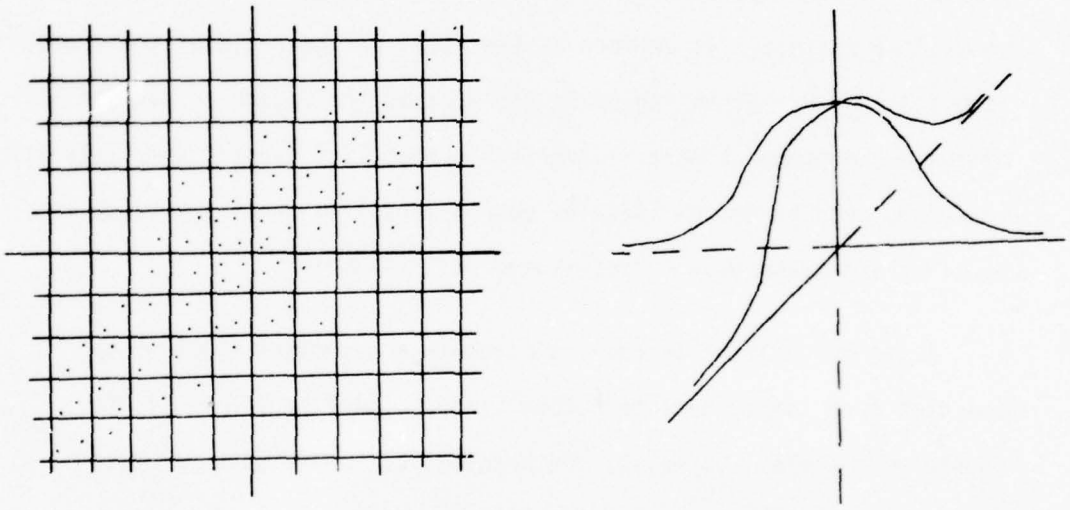
In choosing the level of significance for both the mean and variance tests, one must bear in mind that minimizing the probability

of rejecting the equality hypothesis when it is, in fact, true increases the probability of the opposite kind of error, accepting the equality hypothesis when it is, in fact, false (that is, when the two means or variances are not equal). If they are not equal they must differ by some amount, the magnitude of which strongly affects the power of the test which is the probability of rejecting the equality hypothesis when it is false. Clearly when two statistics are not equal it is much easier to reject an equality hypothesis if the difference is large rather than small. In such circumstances the level of significance should be set at the highest acceptable value in order to maximize the power of detecting a given difference or minimizing the difference detected with a given power.

SECTION V

STATISTICAL MEASURES OF INTERDEPENDENCE AMONG VARIABLES

Multiple measurements for two random variables can be characterized by a measure of average value and dispersion for each variable plus some measure of the interdependence or correlation between the two variables. In Figure 12a, a bivariate observation is represented by the coordinates of each point and a subdivision of both coordinate axes is represented by each grid square.



a. Scatter Plot

b. Bivariate Frequency Distribution Function

Figure 12. Multiple Observations for Two Random Variables

A bivariate histogram analogous to Figure 1a could be constructed from Figure 12a by erecting above each grid square a bar with height equal to the number of observations. For large numbers of observations and more refined subdivisions, such bivariate histograms

approximate continuous bivariate mathematical functions, $f(x,y)$, illustrated by the surface in Figure 12b. The total volume under the surface $f(x,y)$ represents the total number of observations and the volume between the surface and any region in the XY plane represents the expected number of observations in that region. If the volume under the surface is normalized to one by dividing $f(x,y)$ by the total number of observations, then the volume between the surface and any region in the xy plane represents the probability that a randomly selected bivariate observation will fall in that region. That bivariate probability density, $p(x,y)$, the value of the ordinate for such a normalized surface, is defined as the limit of the ratio of the probability associated with a given region to the area of the region as each of its dimensions approaches zero. Bivariate probability density functions are therefore used to define infinite populations from which any given set of bivariate measurements constitutes only a sample.

As in the univariate case, statistics are numbers selected or computed from sample data to define various characteristics of the sample as a whole. Likewise, the mathematical expression for every bivariate probability density function contains one or more constant parameters that define characteristics of the whole population. Sample statistics are therefore used to estimate unknown population parameters describing the same general characteristic. For each of the two variables in a set of bivariate data, the characteristics of

chief interest are the same measures of mean value, dispersion, skewness, and kurtosis that were defined for the univariate case. One additional statistic for the bivariate case defines the degree of interdependence or correlation between the two variables.

In Figure 12a, the standard deviations of the x and y coordinates are respectively the horizontal and vertical root mean square deviations of each point from an axis system with its origin at the mean value coordinates. For this axis system a positive correlation between the two variables is indicated by a predominance of observation points in the first and third quadrants, while a preponderance of observation points in the second and fourth quadrants would indicate a negative correlation. This property is quantified in measures of correlation defined in the following paragraphs.

1. MEASURES OF BIVARIATE CORRELATION

a. Coefficient of Correlation

A measure of the degree of relationship or association between two variables is the coefficient of correlation given by summing the products of the standardized values for each observation and dividing by the number of observations. Expressed mathematically

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-y}{s_y}\right)$$

$$\rho = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)p(x,y)\,dy\,dx \tag{22}$$

Replacing s and σ by their definitions yields

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\rho = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)\,p(x,y)\,dy\,dx}{\sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)^2\,p(x,y)\,dy\,dx}\sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_y)^2\,p(x,y)\,dy\,dx}} \tag{22a}$$

Multiplying products then summing and dividing by n yields

$$r = \frac{\overline{xy}-\bar{x}\bar{y}}{\sqrt{\overline{x^2}-\bar{x}^2}\sqrt{\overline{y^2}-\bar{y}^2}}$$

$$\rho = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\,p(x,y)\,dy\,dx-\mu_x\mu_y}{\sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2p(x,y)\,dy\,dx-\mu_x^2}\sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}y^2p(x,y)\,dy\,dx-\mu_y^2}} \tag{22b}$$

In any of these alternative definitions setting $y = x$ makes $r = 1$ and $\rho = 1$, setting $y = -x$ makes $r = -1$ and $\rho = -1$. Thus a correlation coefficient of one (or minus one) implies perfect positive (or negative) correlation; that is, each bivariate measurement is the same number of its standard deviation units away from its mean value, in the same direction for positive correlation and in the opposite direction for negative correlation. Perfectly correlated variables are therefore identical except for possible differences in the reference point and scaling unit, as for example temperature in °F or °C. Zero correlation implies no relation between the two variables since, in this case, the sums in the numerators of Equations 22 and 22a above contain offsetting positive and negative contributions.

In defining the standard deviation an unbiased estimate for the population was obtained by dividing the sum of the squared deviations from the mean by n-1 rather than n whenever that mean was computed from the same sample data as the standard deviation. A similar consideration exists for the sample correlation. In this case an unbiased estimate of the correlation ρ in a bivariate population is given as

$$\rho = \sqrt{[(n - 1)\, r^2 - 1]/(n - 2)} \qquad (22c)$$

ρ = correlation estimate for a bivariate population

r = correlation computed from bivariate samples

n = number of samples

b. Autocorrelation (Serial Correlation)

If, in the defining equations for correlation (22, 22a and 22b), the observations, x_i, are measurements taken sequentially during n equal intervals of time, Δt, and the observations, y_i, are values not of a second variable but of the first variable measured m intervals of time later, $y_i = x_{i+m}$ with $m \ll n$, then the coefficient computed is the autocorrelation for time lag $\tau = m\Delta t$. If we use this notation the Expressions 22a and 22b for the autocorrelation become

$$r_a = \frac{\sum_{i=1}^{n}\left[x(t_i)-\bar{x}\right]\left[x(t_i+m)-\bar{x}\right]}{\sum_{i=1}^{n}\left[x(t)-\bar{x}\right]^2}$$

$$\rho_a = \frac{\lim_{T\to\infty}\frac{1}{T}\int_0^T\left[x(t)-\bar{x}\right]\left[x(t+\tau)-\bar{x}\right]dt}{\lim_{T\to\infty}\frac{1}{T}\int_0^T\left[x(t)-\bar{x}\right]^2 dt} \qquad (23)$$

$$r_a = \frac{\overline{x(t)\,x(t+\tau)}-\bar{x}^2}{\overline{x^2}-\bar{x}^2}$$

$$\rho_a = \frac{\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)\,x(t+\tau)\,dt-\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt\right]^2}{\lim_{T\to\infty}\frac{1}{T}\int_0^T x^2(t)\,dt-\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt\right]^2} \qquad (23a)$$

In books and journals on time series statistics these expressions are called normalized autocorrelations, with the autocorrelation itself referring only to the first term of the numerator of Equation 23a. For both Equations 23 and 23a the full numerator is the autocovariance function and the denominator is, of course, simply the variance of the sequence of measurements. For some applications the intervals of time in these definitions may be replaced by intervals of distance along some pathway in space.

By computing autocorrelations for several sequential values of m (and thus of τ since $\tau = m\Delta t$) an autocorrelation function is obtained. Autocorrelation functions are useful in identifying periodicities in sequential statistical data. For example, hourly measurements of temperature have a diurnal cycle normally rising from a dawn low to an afternoon high and then falling again. Consequently the expected autocorrelation function would be cyclical with maximum positive correlation at multiples of $m\Delta t$ = 24 hours and negative correlations at the half multiples in between. For very long records this daily cycle would be superimposed upon a similar annual cycle from winter lows to summer highs.

c. Cross-Correlation

If in the defining equations for correlation (22, 22a, 22b), the observations x_i are again measurements taken sequentially during n equal intervals of time Δt, and the observations y_i are values of a second variable measured m intervals of time later with $m << n$, then the coefficient computed is the cross-correlation for time lag $\tau = m\Delta t$. If we use this notation the Expressions 22a and 22b for cross-correlation become

$$r_C = \frac{\sum_{i=1}^{n}\left[x(t_i)-\bar{x}\right]\left[y(t_i+m)-\bar{y}\right]}{\sqrt{\sum_{i=1}^{n}\left[x(t_i)-\bar{x}\right]^2}\sqrt{\sum_{i=1}^{n}\left[y(t_i+m)-\bar{y}\right]^2}} \tag{24}$$

$$\rho_C = \frac{\lim_{T\to\infty}\frac{1}{T}\int_0^T\left[x(t)-\bar{x}\right]\left[y(t+\tau)-\bar{y}\right]dt}{\sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T\left[x(t)-\bar{x}\right]^2 dt}\sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T\left[y(t+\tau)-\bar{y}\right]^2 dt}}$$

$$r_C = \frac{\overline{x(t)y(t+\tau)}-\bar{x}\bar{y}}{\sqrt{\overline{x^2}-\bar{x}^2}\sqrt{\overline{y^2}-\bar{y}^2}} \tag{24a}$$

$$\rho_C = \frac{\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)y(t+\tau)dt-\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt\right]\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T y(t)dt\right]}{\sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T x^2(t)-\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T x(t)dt\right]^2}\sqrt{\lim_{T\to\infty}\frac{1}{T}\int_0^T y^2(t)dt-\left[\lim_{T\to\infty}\frac{1}{T}\int_0^T y(t)dt\right]^2}}$$

Note that cross-correlation becomes simple correlation for $m = 0$ or $\tau = 0$ and it becomes autocorrelation for $x(t + \tau) = y(t + \tau)$. Accordingly, some statements similar to those for autocorrelation can be made. In the literature about time series statistics, the above expressions are called normalized cross-correlations with the cross-correlation itself referring only to the first term of the numerator of Equation 24a. For both Equations 24 and 24a the full numerator is the cross-covariance function and the denominator is, of course, simply the product of the standard deviations for the two measurement sequences. For some applications the intervals of time in these definitions may be replaced by intervals of distance along some pathway in space.

By computing cross-correlations for several sequential values of m (and thus of τ since $\tau = m\Delta t$) a cross-correlation function is obtained. Cross-correlation functions are useful in measuring response times to some prior stimulation or excitation. For example, if a system disturbance originates at point A and is transmitted directly to point B in time τ_1 and indirectly to the same point in time τ_2, then the cross-correlation function would be expected to rise to a strong relative maximum at time τ_1, and a weaker relative maximum at τ_2. Conversely, knowledge of such response times from cross-correlation maxima can be useful in identifying transmission paths of disturbances in complex systems.

d. Rank Correlation

If in the defining equations for correlation (22, 22a, 22b) the observations x_i, y_i are not measurements of continuously variable magnitudes but represent instead the rank order of observation i among all observations for the same variable, then the coefficient computed is the rank correlation. In other words, if observations of two variables are both independently ranked from lowest to highest and these ranks rather than any measured values are used for the x_i and y_i in Equations 22 and 22a, then the resulting value is the coefficient of rank correlation. It is equivalent to the following alternate form derived in Appendix A

$$r = 1 - 6 \sum d_i^2/n(n^2 - 1) \qquad \text{where} \qquad (25)$$

d_i = rank difference for observation pair i

n = sample size

For y = x, the ranks of the two variables are identical for all observations, making $d_i = 0$ for all i and r = 1. For y = -x the two variables have inverted rank orders and r = -1. Thus perfect positive and perfect negative correlation are indicated respectively by plus one and minus one coefficients of rank correlation. Zero correlation is associated with random rank pairings that characterize two unrelated variables.

Rank correlation is most appropriate for variables which are spoken of in quantitative terms but are not capable of objective measurement. System complexity, for example, is a composite variable which cannot be measured in any single unit but it is sufficiently well understood that several systems can generally be ranked in order of their complexity.

e. Point Biserial Correlation

If in the defining equations for correlation (22, 22a, 22b) one of the variables, say x, is not the measurement of a continuous magnitude but is a discrete variable limited to either zero or one values, then the coefficient computed is called the point biserial correlation. It is equivalent to the following alternate form derived in Appendix B.

$$r_p = \frac{\bar{y}_1 - \bar{y}_0}{s_y} \sqrt{p_0 p_1} \tag{26}$$

$\bar{y}_0$ = mean of y measurements for which x = 0

$\bar{y}_1$ = mean of y measurements for which x = 1

s_y = standard deviation of all y measurements

p_0 = proportion of y measurements for which x = 0

p_1 = proportion of y measurements for which x = 1

If the y means of the two groups are equal, the point biserial correlation is zero, whatever the values of p_0, p_i, and s_y. With equal proportions the point biserial correlation can equal one only if the difference in y means is twice s_y.

Point biserial correlation is appropriate for paired observations one of which is a continuous measurement and the other is a simple dichotomous measurement that classifies each observation into one of two categories. The classification may be a quantitative one in which the one or zero represents the presence or absence of a particular attribute, or it may be a purely qualitative one in which the zero/one assignment is made arbitrarily.

f. Tetrachoric Correlation

If in the defining equations for correlation (22, 22a, 22b) both the x and y observations are not continuous magnitudes but discrete variables limited to either zero or one values, then the coefficient computed is called the tetrachoric correlation.* It is equivalent to the following alternate form derived in Appendix C.

$$r_{\phi} = \frac{ad-bc}{\sqrt{(a + c)\ (b + d)}\sqrt{(a + b)\ (c + d)}} \tag{27}$$

a = number of observations for which x = 0, y = 0

b = number of observations for which x = 1, y = 0

c = number of observations for which x = 0, y = 1

d = number of observations for which x = 1, y = 1

Clearly, if b = c = 0 then r = 1; if a = d = 0, then r = -1; and if ad = bc then r = 0.

Tetrachoric correlation is appropriate for paired observations each of which is a simple dichotomous classification of a single observation into one of two categories. As before, the classification may be a quantitative one in which the one or zero represents the presence or absence of a particular attribute, or it may be a purely qualitative one in which the zero/one assignment is made arbitrarily.

* - The term tetrachoric correlation is sometimes reserved for the case in which both x and y are continuous variates arbitrarily reduced to two categories above and below some selected level.

The square of the tetrachoric correlation is related to chi-square computed from the same data:

$$r_\phi^2 = \chi^2/n \qquad \text{where}$$

$$\chi^2 = \frac{[a - (a + b)\,(a + c)/n]^2}{(a + b)\,(a + c)/n} + \frac{[b - (b + a)\,(b + d)/n]^2}{(b + a)\,(b + d)/n} + \frac{[c - (c + a)\,(c + d)/n]^2}{(c + a)\,(c + d)/n} + \frac{[d - (d + b)\,(d + c)/n]^2}{(d + b)\,(d + c)/n} \tag{27a}$$

The numerator of each of the four terms on the right is the square of the difference between the actual number of observations and the expected number assuming the effect of changes in one of the variables is independent of the value of the other. That is to say, there is no interdependence or correlation between them.

g. Contingency Tables

From the defining equations for correlation (22, 22a, 22b) point biserial and tetrachoric correlations were obtained by assigning zero and one values to the variable representing the dichotomous classification of the statistical observations. If there are more than two classes in either or both variables then this procedure cannot be used. However, for multichotomous classifications of both variables, it is still possible to compute the chi-square quantity given for the two by two case in Equation 27a.

Consider for example, the three by four contingency table given below:

a	b	c	d	(a+b+c+d)
e	f	g	h	(e+f+g+h)
i	j	k	l	(i+j+k+l)
(a+e+i)	(b+f+j)	(c+g+k)	(d+h+l)	

The actual number of observations in the row one, column one position is a. The expected number assuming no interdependence is that number a_o which bears the same ratio to the total number of observations in column one that the total number of observations in row one bears to the total number of observations in the table. Expressed mathematically

$$\frac{a_o}{a + e + i} = \frac{a + b + c + d}{(a + b + c + d) + (e + f + g + h) + (i + j + k + l)}$$

The remaining expected values are computed by a corresponding use of other marginal totals. Chi-square for this case is then given by

$$\chi^2 = \frac{(a - a_o)^2}{a_o} + \frac{(b - b_o)^2}{b_o} + \frac{(c - c_o)^2}{c_o} + \frac{(d - d_o)^2}{d_o}$$
$$+ \frac{(e - e_o)^2}{e_o} + \frac{(f - f_o)^2}{f_o} + \frac{(g - g_o)^2}{g_o} + \frac{(h - h_o)^2}{h_o}$$
$$+ \frac{(i - i_o)^2}{i_o} + \frac{(j - j_o)^2}{j_o} + \frac{(k - k_o)^2}{k_o} + \frac{(l - l_o)^2}{l_o}$$

The number of degrees of freedom for two way contingency tables is the product of the number of rows minus one and the number of columns minus one. For the example this is (3 - 1) x (4 - 1) = 2 x 3 = 6. If this computed χ^2 exceeds the tabulated χ^2 for the same number of degrees of freedom then a hypothesis of no relationship or inter-dependence between the two classifications can be rejected with a probability of error not greater than the level of significance for the χ^2 table employed.

h. Correlation of Attributes

Contingency table classifications describing characteristics of objects, systems, or processes are referred to as attributes. The degree of relationship, association, or interdependence among the classifications in a k by k contingency table is called the correlation of attributes, r_ϕ, given by the expression

$$r_\phi = \sqrt{\chi^2/n(k - 1)} \tag{28}$$

where χ^2 and n are as previously defined. For k = 2, Equation 28 defines tetrachoric correlation, thus accounting for the identical ϕ subscript found in Equation 27a. Identical row distributions in all columns and identical column distributions in all rows implies zero correlation. All observations on the diagonal implies a perfect correlation of one. Diagonals do not exist in nonsquare contingency tables, so an alternate measure of independence having less than unit maximum value is defined in the following paragraph.

i. Coefficient of Contingency

A measure of the degree of relationship, association, or interdependence of the classifications in an r by c contingency table is the coefficient of contingency given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{1 - \left[1 \Big/ \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{f_{ij}}{r_i \, c_j}\right]} \qquad (29)$$

f_{ij} = number of observations in row i column j of the contingency table

r_i = number of observations in row i

c_j = number of observations in column j

$n = \sum_i r_i = \sum_j c_j = \sum_i \sum_j f_{ij}$ = total number of observations

χ^2 = the sum of ratios as described in previous paragraph "g"

The equality of these two expressions for the coefficient of contingency C is established in Appendix D. If equality exists in each position of a contingency table between the actual number of observations and the expected number assuming no interdependence, then both χ^2 and C will equal zero. The larger the value of C, the greater is the degree of interdependence. If quantitative bivariate measurements are subdivided into a large number of interval categories for each variate and each observation classified into the resulting contingency table, then the coefficient of contingency for the categorized observations approaches the coefficient of correlation for the quantitative measurements. The number of rows and columns in a contingency table determines the maximum value of C, which is given by $\sqrt{(k-1)/k}$ for the case in which there are k rows and k columns. For this special case only, an alternate measure of interdependence having a maximum value of one for perfect correlation is the correlation of attributes given in paragraph h.

2. MEASURES OF MULTIVARIATE CORRELATION

a. Multiple Correlation

A measure of the degree of relationship or association between one variable and two or more others taken together is the coefficient of multiple correlation. For the simplest case of three variables this is given in terms of the bivariate correlations by the expression

$$r^2_{0\cdot 12} = \frac{r^2_{01} + r^2_{02} - 2r_{01}\, r_{02}\, r_{12}}{1 - r^2_{12}} \qquad (30)$$

$r_{0\cdot 12}$ = multiple correlation between variable 0 and variables 1 and 2

r_{01} = correlation between variables 0 and 1

r_{02} = correlation between variables 0 and 2

r_{12} = correlation between variables 1 and 2

This coefficient of multiple correlation is equivalent to the simple bivariate correlation between the measured values of a dependent variable x_{0i} and their corresponding estimates computed from a linear combination of the independent variables (x_{1i}, x_{2i}), $x'_{0i} = b_0 + b_1 x_{1i} + b_2 x_{2i}$, where the coefficients b_0, b_1, and b_2 are chosen to minimize the mean square error $\sum_{i=1}^{n} (x_{0i} - x'_{0i})^2/n$. The procedure for determining the b values according to this criterion is given by regression theory.

Three special cases of Equation 30 are of interest: $r_{12} = 0$, r_{02} (or r_{01}) $= 0$, and $r_{01} = r_{02}$. If $r_{12} = 0$ then $r^2_{0\cdot 12} = r^2_{01} + r^2_{02}$.

In words, if there is no correlation between the independent variables the square of the multiple correlation between the dependent variable and both independent variables is equal to the sum of the squares of the simple bivariate correlations between the dependent variable and each independent variable.

If $r_{02} = 0$ then $r^2_{0\cdot12} = r^2_{01}/(1 - r^2_{12})$. In this case as r^2_{12} increases from 0 to 1, $r^2_{0\cdot12}$ increases from r^2_{01} to 1. This is surprising. Since there is no correlation at all between variables zero and two, it might be expected that the simple correlation between variables zero and one would equal the multiple correlation between variable zero and both one and two. However, this is true only if there is also no correlation between variables one and two. If they are correlated then variations in variable two will produce variations in variable one but not in variable zero (since $r_{02} = 0$). Thus with variable two accounting for some of the variations in variable one the remaining variation is more closely associated with variations in variable zero than the simple bivariate correlation between them would indicate. Therefore, increasing positive or negative correlation between two independent variables one of which has no correlation with the dependent variable produces an increasing multiple correlation between the dependent and independent variables.

Turning now to the third special case of Equation 30, if $r_{01} = r_{02} = r_0$, then $r^2_{0\cdot12} = 2r^2_0/(1 + r_{12})$. In this instance as r_{12} increases from 0 to 1, $r^2_{0\cdot12}$ decreases from $2r^2_0$ to r^2_0; but as r_{12} decreases from 0 to -1, $r^2_{0\cdot12}$ increases from $2r^2_0$ to +1. The drop

in $r^2_{0\cdot12}$ as r_{12} is increasingly positive results from the fact that both variables one and two are progressively accounting for more and more of the same variations in variable zero, and one of them is therefore becoming more and more redundant. The rise in $r^2_{0\cdot12}$ as r_{12} becomes more negative results from the fact that variables one and two are progressively accounting for more and more of the opposite variations in variable zero, and both of them are therefore becoming more and more critical. Therefore, decreasingly positive or increasingly negative correlation between two independent variables, both of which are positively (or both negatively) correlated with the dependent variable, produces an increasing multiple correlation between the dependent and independent variables. Also, increasingly positive correlation between two independent variables, one of which is positively and the other negatively correlated with the dependent variable, produces an increasing multiple correlation between the dependent and independent variables.

Multiple correlations greater than one will be obtained if certain arbitrary combinations of simple bivariate correlation coefficients are used in Equation 30 and its special cases. This will not occur in practice, however, since intercorrelations among sets of variables cannot be chosen arbitrarily subject only to the condition that they are equal to or less than one in absolute value. Two variables, for example, which are perfectly correlated with a third variable must of necessity be perfectly correlated with each

other. Given correlations among three variables r_{01}, r_{02} and r_{12} it can be shown that the limits for correlation r_{12} will always be

$$r_{12} = r_{01}\, r_{02} \pm \sqrt{1 - r_{01}^2 - r_{02}^2 + r_{01}^2\, r_{02}^2} \tag{31}$$

Extension of multiple correlation to more than three variables is quite simple if matrix notation is employed. One need only define the correlation matrix R to be the array of all the simple bivariate correlations among the complete set of variables.

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots\cdots & r_{1n} \\ r_{12} & 1 & r_{23} & \cdots\cdots & r_{2n} \\ r_{13} & r_{23} & 1 & \cdots\cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{1n} & r_{2n} & r_{3n} & \cdots\cdots & 1 \end{bmatrix} \tag{32}$$

r_{ij} = element in row i col j and row j col i of R

r^{ij} = element in row i col j and row j col i of R^{-1} where R^{-1} is the inverse of matrix R defined so that $RR^{-1} = R^{-1}R = I$, I being the identity matrix having ones in the principal diagonal from upper left to lower right with zeros in all other positions. (33)

The multiple correlation between variable i and all others is then given by

$$r_{i \cdot 1,2,\ldots i-1,i+1,\ldots n} = \sqrt{1 - \frac{1}{r_{ii}\, r^{ii}}} = \sqrt{1 - \frac{1}{r^{ii}}} \tag{34}$$

since $r_{ii} = 1$ for a correlation matrix.

In defining the coefficient of correlation an expression, Equation 22c, was given for obtaining an unbiased estimate of correlation in a bivariate population from the value computed from sample measurements. Similarly Equation 35 below gives an unbiased estimate of the population multiple correlation ρ between variable i and all other variables from the value r computed from sample measurements

$$\rho_{i\cdot 1,2\ldots i-1,i+1,\ldots m} = \sqrt{1 - \left(\frac{n-1}{n-m}\right)\left(1 - r^2_{i\cdot 1,2,\ldots i-1,i+1,\ldots m}\right)} \quad (35)$$

ρ = multiple correlation estimate for population

r = multiple correlation computed from samples

m = number of variables

n = number of multivariate observations

Clearly with m = 2 this reduces to the previous Equation 22c.

b. Marginal Correlation

The multiple correlation between one variable and some of the remaining variables with the rest of the remaining variables ignored is called a marginal correlation. The simple bivariate correlation is a marginal correlation with all but two variables ignored.

c. Conditional or Partial Correlation

If two variables are both correlated with a third variable, then observations resulting solely from variations in this third variable will introduce a spurious correlation between the first two variables. A measure of the correlation between two variables that is independent of variations in other correlated variables is called

conditional or partial correlation. For the simplest case of three variables this is given in terms of the simple bivariate correlations by the expression

$$r_{1\cdot 2|3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)}\ \sqrt{(1 - r_{23}^2)}} \qquad (36)$$

$r_{1\cdot 2|3}$ = conditional correlation between variables 1 and 2 for fixed values of variable 3

r_{12} = correlation between variables 1 and 2

r_{13} = correlation between variables 1 and 3

r_{23} = correlation between variables 2 and 3

Multiple conditional correlations may also be defined. To do this, first partition the correlation matrix to separate the conditioned and conditioning variables as follows:

$$R = \left[\begin{array}{ccc|ccc} r_{11} & \cdots & r_{1m} & r_{1,m+1} & \cdots & r_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ r_{m1} & \cdots & r_{mm} & r_{m,m+1} & \cdots & r_{mn} \\ \hline r_{m+1,1} & \cdots & r_{m+1,m} & r_{m+1,m+1} & \cdots & r_{m+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ r_{n,1} & \cdots & r_{n,m} & r_{n,m+1} & \cdots & r_{nn} \end{array}\right] = \left[\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right] \qquad (37)$$

where all diagonal elements r_{ii} equal one and all $r_{ij} = r_{ji}$.

Then compute the matrix T given by

$$T = R_{11} - R_{12} R_{22}^{-1} R_{21} \tag{38}$$

The conditional correlation between variables i and j of the first m variables for fixed values of the last (n-m) variables is then given from elements t_{ij} of matrix T by

$$t_{i \cdot j | m+1,\ldots n} = t_{ij} / \sqrt{t_{ii}} \sqrt{t_{jj}} \tag{39}$$

Multiple conditional correlations are then given by using the matrix of these conditional correlations rather than R in Equation 34:

$$t_{i \cdot 1,2,\ldots i-1,\ i+1,\ldots m | m+1,\ldots n} = \sqrt{1 - \frac{1}{t^{i \cdot i | m+1,\ldots n}}} \tag{40}$$

since $t_{i \cdot i | m+1,\ldots n} = 1$ for any correlation matrix.

d. Canonical Correlation

Correlations between linear combinations of two sets of variables subject to certain restrictions on the coefficients prescribing these linear combinations are called canonical correlations. Specifically suppose there is a p variate population $x_1 x_2 \ldots x_p$ and a q variate population $y_1 y_2 \ldots y_q$ with $p \geq q$ for definiteness. Then p linear combinations of the x's, $u_1, u_2, \ldots u_p$ and q linear combinations of the y's, $v_1, v_2, \ldots v_q$ can be found all with zero mean and unit variance and with covariance $(u_i, u_j) = 0$, covariance $(v_i, v_j) = 0$, and covariance $(u_i, v_j) = 0$ for all $i \neq j$. The correlation between u_i and v_i is then a canonical correlation, at most q of which are non-zero. If u and v are the linear combinations corresponding to the largest canonical correlation, then v is the linear combination of the y's which can be predicted from the x's with the least residual variance, and u is the appropriate linear prediction function.

To obtain canonical correlations and their associated coefficients first partition the simple bivariate correlation matrix to identify the correlations within and between the two sets of variables

$$R = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{bmatrix}$$

Then formulate the eigenvalue problem

$$|R_{yx} R_{xx}^{-1} R_{xy} - \lambda R_{yy}| = 0 \qquad (41)$$

and solve for the eigenvalues $\lambda_1, \ldots, \lambda_q$ and the eigenvectors $\beta_1, \ldots, \beta_q$. The eigenvectors normalized so that $\beta_i R_{yy} \beta_i = 1$ are the canonical coefficients for the standardized y variables. The canonical correlations γ_i and the coefficients α of the standardized x variables are given by

$$\gamma_i^2 = \lambda_i \qquad \alpha = \gamma_i^{-1} R_{xx}^{-1} R_{xy} \beta_i \qquad (41a)$$

e. Autocorrelation

In the previous section on bivariate measures, autocorrelation was defined in terms of a single sequence of time data points. Here autocorrelation is defined in terms of an ensemble of such time history data. In this multiple record case each observation time is a separate variable with an observed measurement from each record. With autocorrelation now defined as the first term in the numerator of Equation 23a, Figure 13 shows both the time correlation for each individual record and the ensemble correlation for each pair of observation times. For a stationary random process using the notation of Figure 13

$$\left\langle y(t_i) \right\rangle = \left\langle y(t_j) \right\rangle \qquad \left\langle y^2(t_i) \right\rangle = \left\langle y^2(t_j) \right\rangle$$
$$\left\langle y(t_i) \quad y(t_j) \right\rangle = \left\langle y(t)\, y(t+\tau) \right\rangle \quad \tau = t_j - t_i \tag{42}$$

For an ergodic random process it is also true that

$$\left\langle y(t) \right\rangle = \bar{y}_i \qquad \left\langle y^2(t) \right\rangle = \overline{y_i^2} \qquad i = 1 \ldots . n$$
$$\left\langle y(t)\, y(t+\tau) \right\rangle = \overline{y(t)\, y(t+\tau)} \tag{43}$$

Stationarity thus implies that the mean and mean square ensemble averages are independent of time and that the autocorrelation depends only on the time difference, not on the particular starting or ending times. Ergodicity implies stationarity and the equality of the ensemble average with the time average of each record for the mean, mean square and autocorrelation values. The converse of these statements is not strictly true since stationarity and ergodicity involve

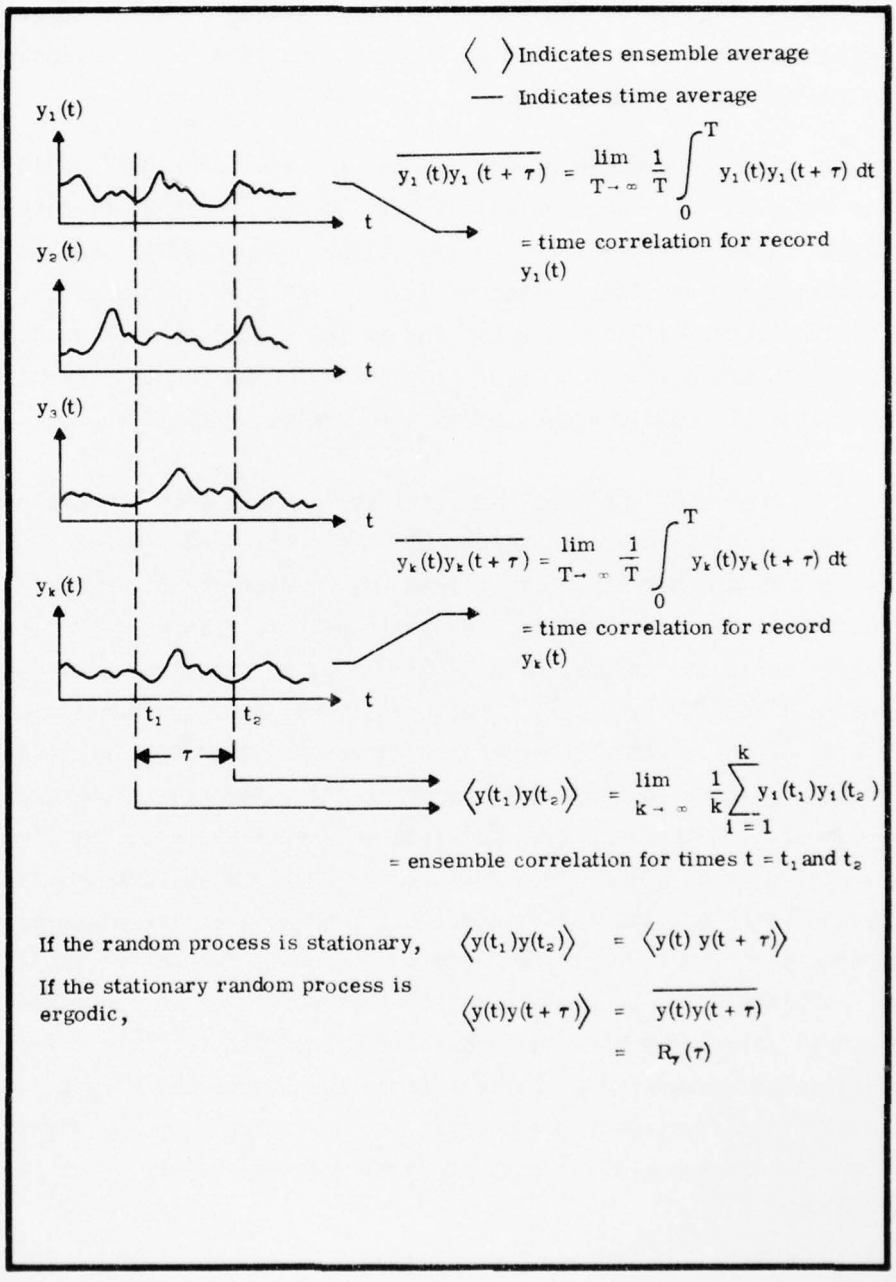


Figure 13. Computation of Autocorrelation Values

all statistics, not just the mean, mean square, and autocorrelation. For practical purposes, however, Equations 42 and 43 are generally considered both necessary and sufficient conditions for stationarity and ergodicity, respectively.

For an ergodic random process the time averaged statistics from any one record are equivalent both to the same time averaged statistics for any other record and to the same ensemble averaged statistics for any time or set of times. Therefore, an analysis of a single record will suffice for the entire ensemble if the random process is ergodic. This is of course the reason for the practical importance of ergodic processes in time series statistics.

For many applications, time series data from ergodic processes are best treated in their reciprocal time or frequency domain. Spectral values of this kind result from the Fourier transform of the random time history function. Multiplying this transform by its own complex conjugate produces the real valued autospectral (or power spectral) density function. The autospectral density function also results directly from the Fourier transform of the autocorrelation function (a real valued transform since the autocorrelation function is symmetric with respect to positive and negative values of the time interval τ in Equations 23 and 23a). The mathematical theory and computational details for carrying out Fourier transformation of time data will not be treated here since a wide literature exists on this subject. For present purposes it is sufficient to note that spectral values may be used in place of time data for the various statistical measures and analysis techniques described herein. As in the bivariate case the time data may be replaced by space data in which measurements are made at equal intervals along a line in space.

f. Cross Correlation

In the previous section on bivariate measures, cross-correlation was defined in terms of two different sequences of time data points. If each sequence represents measurements from two different ergodic random processes, then the relationships between the two records also apply to the two ensembles. In particular the cross-correlation function now defined by the first term in the numerator of Equation 24a is

$$R_{xy}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T x(t)\, y(t+\tau)\, dt \qquad (44)$$

For the special case x = y this defines the autocorrelation function.

As noted in the previous section for many applications time data is converted to the frequency domain by means of Fourier transforms. The Fourier transform of x(t) multiplied by the complex conjugate of the Fourier transform of $y(t + \tau)$ gives the cross spectral density function $G_{ij}(f)$. The cross-spectral density function also results directly from the Fourier transform of the cross-correlation function given by Equation 44.

For more than two ensembles spectral density functions between every pair may be computed and arranged in a matrix with auto-spectra G_{ii} in the diagonal positions and cross-spectra G_{ij} in the off diagonal positions. Corresponding cross-spectral elements on opposite sides of the diagonal G_{ij} and G_{ji} will be complex conjugates of one another. The coherence function $\gamma^2_{ij}(f)$ between record i and j is then given by

$$\gamma^2_{ij}(f) = \frac{G_{ij}(f)\, G_{ji}(f)}{G_{ii}(f)\, G_{jj}(f)} \leq 1 \qquad (45)$$

Upon arranging the γ_{ij} into matrices, various multiple, marginal, conditional, and canonical coherences may be defined in ways analogous to correlations of the same kind.

SECTION VI

FACTOR ANALYSIS OF MULTIPLE VARIABLES

The concept underlying factor analysis is best illustrated by an example. Consider the following correlation matrix among eight variables.

VARIABLE NUMBER	1	2	3	4	5	6	7	8
1	1.0000	.2208	.0624	.8088	.0888	.7800	.2952	.2208
2	.2208	1.0000	.3080	.2972	.2540	.3040	.8520	.1356
3	.0624	.3080	1.0000	.3448	.8912	.1416	.5176	.8024
4	.8088	.2972	.3448	1.0000	.3504	.8948	.4380	.4772
5	.0888	.2540	.8912	.3504	1.0000	.1608	.4472	.7460
6	.7800	.3040	.1416	.8948	.1608	1.0000	.4000	.2872
7	.2952	.8520	.5176	.4380	.4472	.4000	1.0000	.3216
8	.2208	.1356	.8024	.4772	.7460	.2872	.3216	1.0000

First reorder the row and column numbers as follows:

Old Number	1	2	3	4	5	6	7	8
New Number	6	4	1	8	3	7	5	2

Then rewrite the above eight by eight matrix as follows:

VARIABLE NUMBER	3	8	5	2	7	1	6	4
3	1.0000	.8024	.8912	.3080	.5176	.0624	.1416	.3448
8	.8024	1.0000	.7460	.1356	.3216	.2208	.2872	.4772
5	.8912	.7460	1.0000	.2540	.4472	.0888	.1608	.3504
2	.3080	.1356	.2540	1.0000	.8520	.2208	.3040	.2972
7	.5176	.3216	.4472	.8520	1.0000	.2952	.4000	.4380
1	.0624	.2208	.0888	.2208	.2952	1.0000	.7800	.8080
6	.1416	.2872	.1608	.3040	.4000	.7800	1.0000	.8948
4	.3448	.4772	.3504	.2972	.4380	.8080	.8948	1.0000

In this matrix the correlations in the three blocks along the diagonal are all very high while those in the six off diagonal blocks are very low. Thus the variables fall into three groups (3, 8, 5), (2, 7), and (1, 6, 4) characterized by high within group correlations and low between group correlations. Each group therefore represents a factor that is measured rather well by any variable within the group and very poorly by any variable outside the group. Of course, not all correlation matrices can be reordered with such a clear distinction between correlations in the diagonal and off diagonal blocks, but this would only represent cases in which some of the variables are strongly affected by two or more of the factors.

Continuing with this example, one can verify that all of the off diagonal correlations in the reordered correlation matrix can be obtained exactly by the following product of a matrix and its transpose

$$\begin{bmatrix} .96 & .24 & .04 \\ .82 & .02 & .26 \\ .88 & .18 & .08 \\ .10 & .86 & .14 \\ .30 & .92 & .22 \\ .00 & .12 & .84 \\ .06 & .16 & .90 \\ .28 & .20 & .94 \end{bmatrix} \begin{bmatrix} .96 & .82 & .88 & .10 & .30 & .00 & .06 & .28 \\ .24 & .02 & .18 & .86 & .92 & .12 & .20 & .16 \\ .04 & .26 & .08 & .14 & .22 & .84 & .90 & .94 \end{bmatrix}$$

These are the coefficients in an equation representing each variable V_j as a linear combination of the three more fundamental factors F_p:

$V_3 = .96\ F_1 + .24\ F_2 + .04\ F_3$ $\qquad$ $V_2 = .10\ F_1 + .86\ F_2 + .14\ F_3$

$V_8 = .82\ F_1 + .02\ F_2 + .26\ F_3$ $\qquad$ $V_7 = .30\ F_1 + .92\ F_2 + .22\ F_3$

$V_5 = .88\ F_1 + .18\ F_2 + .08\ F_3$

$V_1 = .00\ F_1 + .12\ F_2 + .84\ F_3$

$V_6 = .06\ F_1 + .16\ F_2 + .90\ F_3$

$V_4 = .28\ F_1 + .20\ F_2 + .94\ F_3$

Note that the first group of variables (3, 8, 5) is most heavily loaded on the first factor F_1, the second group (2, 7) on the second factor F_2, and the third group (1, 6, 4) on the third factor F_3. Clearly the sum of the squares of the coefficients in each of these equations is one of the diagonal terms in the above matrix product. This quantity is called the communality and it represents the square of the correlation between each variable and its common factor representation as

given above. The difference between the communality and one represents the effect of a unique factor associated with each variable.

The matrix product shown above is not unique -- a correlation matrix R can be decomposed into many such products. In factor analysis the one having maximum variance among the elements is chosen since this one also maximizes the number of elements having very low and very high absolute values. This, in turn, simplifies the interpretation of each factor by representing each of them in terms of the smallest number of variables.

1. FACTOR ANALYSIS MODEL

The object of factor analysis is to define a large number of interrelated variables in terms of a much smaller number of more independent factors. The simplest mathematical model for describing a variable in terms of several others is the linear representation. For such a linear composite to be valid, however, all variable and factor measurements must be referenced to the same origin and scaled in the same units. To do this, one first subtracts from each observation x_i its mean value $\bar{x}$ and then divides the resultant quantity by its standard deviation s_x, a measure of the dispersion or scatter in a set of observations. Thus transformed, the new standardized value expresses the deviation from the mean in standard deviation units. Expressed mathematically:

$$z_{ji} = (x_{ji} - \bar{x}_j)/s_{xj} \qquad \begin{matrix} i = 1....N \\ j = 1....n \end{matrix} \qquad (46)$$

where

$$\bar{x}_j = \frac{1}{N} \sum_{i=1}^{N} x_{ji} \qquad s_{xj} = \frac{1}{N} \sum_{i=1}^{N} (x_{ji} - \bar{x}_j)^2$$

The classical factor analysis model may be written for the standardized value of the jth variable and the ith observation as follows:

$$z_{ji} = \sum_{p=1}^{m} a_{jp} F_{pi} + d_j U_{ji} \qquad \begin{matrix} j = 1...n \\ i = 1...N \\ m<n \end{matrix} \qquad (47)$$

In this expression F_{pi} is the standardized value of the common factor F_p for observation i, each of the m terms $a_{jp} F_{pi}$ represents the contribution of the corresponding factor to the linear composite, and the $d_j U_{ji}$ is the residual, specific, or unique contribution in the assumed representation of the observed measurement z_{ji}. In the geometric representation of this model, the unique factors are assumed to be mutually orthogonal and orthogonal to the common factors which are not necessarily assumed mutually orthogonal. Note that the representation is not unique since the total number of factors F_p, U_j exceeds the number of variables, z_j.

The complete set of N values for each of the n variables can be represented by the n x N matrix as follows:

$$Z = \begin{bmatrix} z_{11} \cdots\cdots z_{1N} \\ \cdots\cdots\cdots\cdots \\ z_{n1} \cdots\cdots z_{nN} \end{bmatrix}$$

Similarly, the common and unique factors may be represented as

$$F = \begin{bmatrix} F_{11} \cdots\cdots F_{1N} \\ \cdots\cdots\cdots\cdots \\ F_{m1} \cdots\cdots F_{mN} \end{bmatrix} \qquad U = \begin{bmatrix} U_{11} \cdots\cdots U_{1N} \\ \cdots\cdots\cdots\cdots \\ U_{n1} \cdots\cdots U_{nN} \end{bmatrix}$$

The coefficients of these factors in Equations 47 may be represented by the n by m and n by n matrices as follows:

$$A = \begin{bmatrix} a_{11} \cdots\cdots a_{1m} \\ \cdots\cdots\cdots\cdots \\ a_{n1} \cdots\cdots a_{nm} \end{bmatrix} \qquad D = \begin{bmatrix} d_1 \cdots\cdots 0 \\ \cdots\cdots\cdots \\ 0 \cdots\cdots d_n \end{bmatrix}$$

With these definitions, Equation 47 may be written in matrix form

$$Z = AF + DU \tag{48}$$

The matrix of observed correlations among the variables can be defined in matrix notation by

$$R = ZZ'/N \qquad Z' = Z \text{ transpose} \tag{49}$$

If Equation 48, the factor analysis model for the matrix Z, is substituted into this expression, we have

$$R = (AF + DU)\,(AF + DU)'/N$$

$$R = A(FF'/N)\,A' + A(FU'/N)D'$$

$$+ D(UF'/N)\,A' + D(UU'/N)D'$$

The first and last quantities in parentheses both having the same form as Equation 49 are correlation matrices. The correlation matrix of the common factors is denoted by $\Phi = FF'/N$. The correlation matrix of the unique factors is an identity matrix since the unique factors are assumed to be uncorrelated, i.e., represented by mutually orthogonal axes. The remaining two terms in parentheses are both null matrices since the common and unique factors are assumed to be uncorrelated, i.e., mutually orthogonal. Thus, we have

$$R = A(FF'/N)A' + D(UU'/N)D = A\,\Phi\,A' + DD' \tag{50}$$

If the common factors are also assumed to be uncorrelated or orthogonal

$$R = AA' + DD' \tag{51}$$

Clearly, the correlation matrix derived from the common factors only is given by

$$R^* = AA' = \begin{bmatrix} a_{11} \cdots\cdots a_{1m} \\ \cdots\cdots\cdots\cdots \\ a_{n1} \cdots\cdots a_{nm} \end{bmatrix} \quad \begin{bmatrix} a_{11} \cdots\cdots a_{n1} \\ \cdots\cdots\cdots\cdots \\ a_{1m} \cdots\cdots a_{nm} \end{bmatrix} \tag{52}$$

This matrix (Equation 52) is the same as the former (Equation 51) in the off-diagonal elements, but the diagonal elements, designated communalities, are numbers less than one. In terms of the matrix elements, they are given by

$$h_j^2 = \sum_{p=1}^{m} a_{jp}^2 \qquad j = 1.....n \tag{53}$$

These communalities are the squares of the correlations $r_{z_j z_1}$ between the total factor and the common factor representations of each of the variables as shown by the following:

Given: $z_{ji} = a_{j1}F_{1i} + \ldots + a_{jm}F_{mi} + d_j U_{j1}$

$z'_{ji} = a_{ji}F_{1i} + \ldots + a_{jm}F_{mi}$

then

$$r_{z_j z'_j} = \sum_{i=1}^{N} z_{ji} z'_{ji} \Big/ \sqrt{\sum_{i=1}^{N} z_{ji}^2 \sum_{i=1}^{N} z'^2_{ji}}$$

$$r_{z_j z'_j} = h_j^2 \Big/ \sqrt{(1)(h_j^2)} = h_j$$

The off-diagonal elements of Equation 50 or 51 are, of course, the ordinary coefficients of correlation given in terms of the matrix elements for the variables j and k by

$$r_{z_j z_k} = \sum_{p=1}^{m} a_{jp} a_{kp} \tag{54}$$

2. NUMBER OF COMMON FACTORS

From matrix theory, it is known that the rank of AA' cannot exceed the rank of A which in turn cannot exceed its smaller dimension, in this

case the number of columns m. Consequently, although the reproduced correlation matrix R* = AA' has order n equal to the number of variables, its rank cannot exceed m, the number of common factors. Since the number of common factors cannot be less than the rank of the reproduced correlation matrix, the minimum number of common factors must equal the minimum possible rank of the reproduced correlation matrix. Since the correlation matrix reproduced from the common factors differs from that reproduced from all the factors only in the diagonal elements, one of the major problems of factor analysis is to determine by how much the rank of a correlation matrix can be reduced from n by a suitable choice of communalities in the diagonal. The computation of such minimal rank communalities is so formidable even on modern computers that it is not normally attempted. Instead, they are approximated by the squared multiple correlations given by one minus the reciprocals of the corresponding elements in the diagonal of the inverse of the correlation matrix. The squared multiple correlations are known to be lower bounds for true minimal rank communalities and approach the latter as the ratio of the number of factors to the number of variables approaches zero.

3. FACTOR SOLUTION

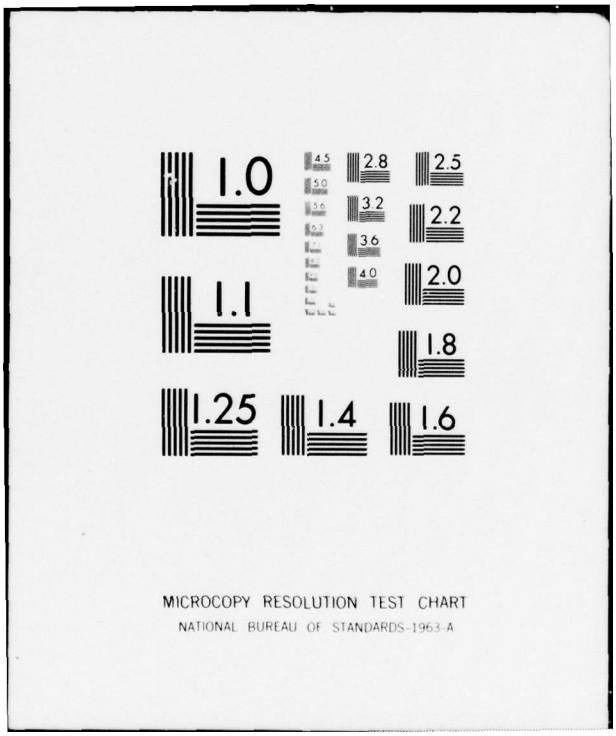
The solution for the a coefficients or loadings in the factor analysis model, Equation 47, is an eigenvalue problem analogous to the one encountered in determining normal modes of vibration or principal

AD-A037 483 AIR FORCE FLIGHT DYNAMICS LAB WRIGHT-PATTERSON AFB OHIO F/G 12/1
STATISTICAL MEASURES, PROBABILITY DENSITIES, AND MATHEMATICAL M--ETC(U)
OCT 76 R G MERKLE

UNCLASSIFIED AFFDL-TR-76-83 NL

2 OF 2
AD
A037483

END
DATE
FILMED
4-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

axes of rotation in dynamics problems. The matrix equation in this case is

$$\begin{bmatrix} h_1^2 - \lambda_p \cdots\cdots r_{1n} \\ \cdots\cdots\cdots\cdots \\ r_{n1} \cdots\cdots\cdots h_n^2 - \lambda_p \end{bmatrix} \begin{bmatrix} a_{1p} \\ \cdots \\ a_{np} \end{bmatrix} = \begin{bmatrix} 0 \\ . \\ 0 \end{bmatrix} \qquad \lambda_p^2 = \sum_{j=1}^{n} a_{jp}^2 \qquad (55)$$

Here the r's are correlation coefficients between the variables, the h's are the communalities or rank minimizing values of the previous section, λ_p is one of the eigenvalues, and the column of a's is the associated eigenvector, the elements of which serve as the coefficients or loadings of the pth factor for the n variables when the $\lambda_p = \sum_{j=1}^{n} a_{jp}^2$ condition is fulfilled. Some of the eigenvalues will be zero since selecting the diagonals to minimize rank is equivalent to minimizing the number of non-zero eigenvalues, or equivalently the number of common factors as desired. When the squared multiple correlation is used to approximate the true rank minimizing communalities in the diagonal, the exact positive semi-definite character of the matrix is destroyed and the zero eigenvalues are replaced with small positive and negative numbers which are simply ignored. In practice, only those factors associated with the few highest eigenvalues are needed in the factor analysis model (Equation 47) since the correlation matrix reproduced from these alone often yields a very close approximation to the observed correlation matrix in the off-diagonal elements which are the elements of consequence.

4. FACTOR ROTATION

The form employed in deriving the factor coefficients or loadings a_{jp} has the property that the sum of the contributions of the successive

factors makes the total communality a maximum under the conditions relating these coefficients to the off-diagonal correlations. As noted previously, no factor solution is unique and other factor loadings not having this property would yield identical correlation matrices.

Since factors are hypothetical constructs, their interpretation must be in terms of the observable variables. The simplest possible illustration of a clear cut factorization occurs when a sequence of variables can be found in which the higher correlations occur in blocks along the principal diagonal of the correlation matrix and the lower correlations occur in all other positions. In terms of the factor analysis model (Equation 47), this corresponds to a number of factors equal to the number of blocks, and each variable having a substantially higher squared loading coefficient on one factor than on any of those remaining, each variable then becoming an imperfect measure of one factor only. In slightly more complex illustrations, maximum squared loadings will occur on several factors with minimum loadings on all those remaining. Ideally then, for the simplest physically meaningful interpretation of the hypothetical factors in terms of observed variables, the squared loading coefficients should approach their upper and lower bounds, one and zero respectively. This implies the maximum possible variance in the squared loading coefficients. Clearly, there must be some orientation of the orthogonal factor axes for which the squared loading coefficients have greater variance than for any other. Mathematically, this requires rotations to maximize the following variance function, the new factor loadings now denoted by b's.

$$\frac{1}{n}\sum_{p=1}^{m}\sum_{j=1}^{n}\frac{b_{jp}^4}{h_j} - \frac{1}{n^2}\sum_{p=1}^{m}\sum_{j=1}^{n}\left[\frac{b_{jp}^2}{h_j^2}\right]^2 \qquad (56)$$

The h's are introduced so that in axis rotations each coefficient is weighted equally rather than in proportion to its communality which would otherwise be the case. The actual rotations required to maximize this function constitute a sequential iteration process. The resulting b_{jp} coefficients are called the varimax loadings.

SECTION VII

MATHEMATICAL MODELS FOR STATISTICAL DATA

Once statistical tests have established that significant differences exist in the means or variances of two or more data samples, there remains the problem of expressing how these differences are related to the different circumstances under which the contrasting sets of measurements were taken. Variations in the circumstances under which data are taken may be expressed quantitatively as in the case of data taken at different temperatures, or the distinctions may be qualitative as in the case of data grouped according to some classification criteria. For statistical predictions, regression functions are employed for quantitative variations in test conditions, analysis of variance models are used for qualitative variations, and analysis of covariance or general linear hypothesis models are formulated to encompass both quantitative and qualitative variations. Only the latter will be treated here since it includes the other two as special cases.

1. MATRIX FORMULATION OF THE MODEL

The form of the mathematical model is most easily seen by generalizing simple cases for one and two way classifications of the qualitative variable each with a single quantitative covariate.

a. One-Way Classification of Variables

For a one-way classification of variable y and covariate x into m groups with n_j measurements in group j:

$$y_{ji} = a + b_j + cx_{ji} + e_{ji}, \quad \sum_{j=1}^{m} b_j = 0, \quad \sum_{i=1}^{n_j} e_{ji} = 0, \quad j=1....m \qquad (57)$$

where

y_{ji} = observation i in group j

a = general value common to all measurements

b_j = an additional contribution characteristic of group j

c = coefficient of covariate x

x_{ji} = covariate value for observation i in group j

e_{ji} = error term for observation i in group j

m = number of groups

n_j = number of observations in group j

For the specific case of three groups with two measurements each, these equations may be written out as follows using the fact that

$$b_3 = -b_1 - b_2 \text{ from } \sum_{j=1}^{3} b_j = 0$$

$$\begin{aligned}
y_{11} &= a + 1\, b_1 + 0\, b_2 + cx_{11} + e_{11} \\
y_{12} &= a + 1\, b_1 + 0\, b_2 + cx_{12} + e_{12} \\
y_{21} &= a + 0\, b_1 + 1\, b_2 + cx_{21} + e_{21} \\
y_{22} &= a + 0\, b_1 + 1\, b_2 + cx_{22} + e_{22} \\
y_{31} &= a - 1\, b_1 - 1\, b_2 + cx_{31} + e_{31} \\
y_{32} &= a - 1\, b_1 - 1\, b_2 + cx_{32} + e_{32}
\end{aligned} \tag{58}$$

In matrix notation this set becomes

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & x_{11} \\ 1 & 1 & 0 & x_{12} \\ 1 & 0 & 1 & x_{21} \\ 1 & 0 & 1 & x_{22} \\ 1 & -1 & -1 & x_{31} \\ 1 & -1 & -1 & x_{32} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \\ c \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix} \tag{59}$$

If one denotes the entire six by one column matrix on the left by Y, the entire six by four matrix on the right by X, the entire four by one column matrix by B, and the entire last six by one column matrix by e, this becomes

$$Y = XB + e \tag{60}$$

Clearly this matrix Equation (60) remains valid for any number of groups and any number of observations per group since only the dimensions of the matrix change. For the same reason additional covariates may be added without affecting the validity of matrix Equation 60.

b. Two-Way Classification of Variables

For a two-way classification of the random variable y and covariate x into m_j rows and m_k columns with n_{jk} measurements in cell jk the analysis of covariance model is

$$y_{jki} = a + b_j + c_k + d_{jk} + fx_{jki} + e_{jki}$$

with

$$\sum_{j=1}^{m_j} b_j = 0, \sum_{k=1}^{m_k} c_k = 0, \sum_{j=1}^{m_j} d_{jk} = 0, \sum_{k=1}^{m_k} d_{jk} = 0 \tag{61}$$

and

$$\sum_{i=1}^{n_{jk}} e_{jki} = 0, \; j = 1 \ldots m_j, \; k = 1 \ldots m_k$$

y_{jki} = observation i in row j column k

a = general value common to all measurements

b_j = main effect for row j

c_k = main effect for column k

d_{jk} = interaction effect, row j column k

f = coefficient of covariate x

x_{jki} = covariate for observation i in row j column k

e_{jki} = error for observation i in row j column k

m_j = number of rows

m_k = number of columns

n_{jk} = number of observation in row j column k

For the specific case of three rows and three columns with two measurements each these equations may be written as follows using the facts that

$b_3 = -b_1 - b_2$, $c_3 = -c_1 - c_2$, $d_{3k} = -d_{1k} - d_{2k}$, and $d_{j3} = -d_{ji} - d_{j2}$

$$y_{111} = a + 1b_1 + 0b_2 + 1c_1 + 0c_2 + 1d_{11} + 0d_{12} + 0d_{21} + 0d_{22} + fx_{111} + e_{111}$$

$$y_{112} = a + 1b_1 + 0b_2 + 1c_1 + 0c_2 + 1d_{11} + 0d_{12} + 0d_{21} + 0d_{22} + fx_{112} + e_{112}$$

$$y_{121} = a + 1b_1 + 0b_2 + 0c_1 + 1c_2 + 0d_{11} + 1d_{12} + 0d_{21} + 0d_{22} + fx_{121} + e_{121}$$

$$y_{122} = a + 1b_1 + 0b_2 + 0c_1 + 1c_2 + 0d_{11} + 1d_{12} + 0d_{21} + 0d_{22} + fx_{122} + e_{122}$$

$$y_{131} = a + 1b_1 + 0b_2 - 1c_1 - 1c_2 - 1d_{11} - 1d_{12} + 0d_{21} + 0d_{22} + fx_{131} + e_{131}$$

$$y_{132} = a + 1b_1 + 0b_2 - 1c_1 - 1c_2 - 1d_{11} - 1d_{12} + 0d_{21} + 0d_{22} + fx_{132} + e_{132}$$

$$y_{211} = a + 0b_1 + 1b_2 + 1c_1 + 0c_2 + 0d_{11} + 0d_{12} + 1d_{21} + 0d_{22} + fx_{211} + e_{211}$$

$$y_{212} = a + 0b_1 + 1b_2 + 1c_1 + 0c_2 + 0d_{11} + 0d_{12} + 1d_{21} + 0d_{22} + fx_{212} + e_{212}$$

$$y_{221} = a + 0b_1 + 1b_2 + 0c_1 + 1c_2 + 0d_{11} + 0d_{12} + 0d_{21} + 1d_{22} + fx_{221} + e_{221}$$

$$y_{222} = a + 0b_1 + 1b_2 + 0c_1 + 1c_2 + 0d_{11} + 0d_{12} + 0d_{21} + 1d_{22} + fx_{222} + e_{222}$$

$$y_{231} = a + 0b_1 + 1b_2 - 1c_1 - 1c_2 + 0d_{11} + 0d_{12} - 1d_{21} - 1d_{22} + fx_{231} + e_{231}$$

$$y_{232} = a + 0b_1 + 1b_2 - 1c_1 - 1c_2 + 0d_{11} + 0d_{12} - 1d_{21} - 1d_{22} + fx_{232} + e_{232}$$

$$y_{311} = a - 1b_1 - 1b_2 + 1c_1 + 0c_2 - 1d_{11} + 0d_{12} - 1d_{21} + 0d_{22} + fx_{311} + e_{311}$$

$$y_{312} = a - 1b_1 - 1b_2 + 1c_1 + 0c_2 - 1d_{11} + 0d_{12} - 1d_{21} + 0d_{22} + fx_{312} + e_{312}$$

$$y_{321} = a - 1b_1 - 1b_2 + 0c_1 + 1c_2 + 0d_{11} - 1d_{12} + 0d_{21} - 1d_{22} + fx_{321} + e_{321}$$

$$y_{322} = a - 1b_1 - 1b_2 + 0c_1 + 1c_2 + 0d_{11} - 1d_{12} + 0d_{21} - 1d_{22} + fx_{322} + e_{322}$$

$$y_{331} = a - 1b_1 - 1b_2 - 1c_1 - 1c_2 + 1d_{11} + 1d_{12} + 1d_{21} + 1d_{22} + fx_{331} + e_{331}$$

$$y_{332} = a - 1b_1 - 1b_2 - 1c_1 - 1c_2 + 1d_{11} + 1d_{12} + 1d_{21} + 1d_{22} + fx_{332} + e_{332}$$

(62)

In matrix notation this becomes

$$
\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \\ y_{311} \\ y_{312} \\ y_{321} \\ y_{322} \\ y_{331} \\ y_{332} \end{bmatrix} =
\begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & x_{111} \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & x_{112} \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & x_{121} \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & x_{122} \\
1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & x_{131} \\
1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & x_{132} \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & x_{211} \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & x_{212} \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & x_{221} \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & x_{222} \\
1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 & x_{231} \\
1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 & x_{232} \\
1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 & x_{311} \\
1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 & x_{312} \\
1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 & x_{321} \\
1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 & x_{322} \\
1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & x_{331} \\
1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & x_{332}
\end{bmatrix}
\begin{bmatrix} a \\ b_1 \\ b_2 \\ c_1 \\ c_2 \\ d_{11} \\ d_{12} \\ d_{21} \\ d_{22} \\ f \end{bmatrix} +
\begin{bmatrix} e_{111} \\ e_{112} \\ e_{121} \\ e_{122} \\ e_{131} \\ e_{132} \\ e_{211} \\ e_{212} \\ e_{221} \\ e_{222} \\ e_{231} \\ e_{232} \\ e_{311} \\ e_{312} \\ e_{321} \\ e_{322} \\ e_{331} \\ e_{332} \end{bmatrix}
\tag{63}
$$

If one denotes the entire eighteen by one column matrix on the left by Y, the entire eighteen by ten matrix on the right by X, the entire ten by one column matrix by B, and the last eighteen by one column matrix by e, this becomes

$$Y = XB + e \tag{60}$$

This is the same matrix equation previously obtained, differing only in having higher dimensions from the addition of a second classification of variables. Clearly the same would apply for three or more classifications of variables. As before the number of groups per classification, the number of measurements per group, and the number of covariates also affect only the dimensions of matrix Equation 60 which remains valid. The elements of matrix B are called regression coefficients.

2. COMPUTING THE MATRIX OF REGRESSION COEFFICIENTS

So far nothing has been said about the means of obtaining the values in the matrix B. Since the mathematical model contains an error term for each measurement, the elements of B could have any values whatsoever and the error term could then be adjusted to make the equality true. The unstated assumption, of course, has been that the errors are to be minimized in some fashion to give a best fit of the function to the data. Minimizing the simple sum of the errors would be inappropriate since large positive and negative errors would offset each other and appear as little or no error. Instead the sum of the squares of the errors is minimized to give the well known least-squares fit.

One of the equalities given by the matrix Equation 60 can be written as

$$y_i = \sum_{j=1}^{m} x_{ji}\, b_j + e_i \tag{64}$$

in which the single subscript i replaces the set of subscripts used above to designate the measurement and the single subscript j replaces the set of subscripts used to distinguish the various group and interaction effects as well as the coefficients of the covariates. The error sum of squares is therefore given by

$$SS_E = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ji}\, b_j\right)^2 \tag{65}$$

To find the values of the b_j for which this is a minimum the derivatives with respect to the b_j are set equal to zero in accordance with the usual procedure in calculus

$$\frac{dSS_E}{db_k} = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ji} b_j\right) x_{ki} = 0 \qquad k=1....m$$

or

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ki} x_{ji} b_j = \sum_{i=1}^{n} x_{ki} y_i \qquad k=1....m$$

This last equation may be written in matrix form

$$X'\, X\, B = X'\, Y$$

These are called the normal equations. They may be solved to obtain the minimizing values of the regression coefficients B:

$$B = (X'\, X)^{-1} X'\, Y \tag{66}$$

3. SIGNIFICANCE TESTS FOR REGRESSION COEFFICIENTS

These elements of B obtained from sample data are only estimates of the true population parameters. Zero values for any subset of these parameters would indicate no contribution to the usefulness of the prediction model (Equation 60). Consequently some statistical test is needed to determine if the computed values for any subset are sufficiently close to zero to warrant such an inference for the population parameters. If this be so the mathematical prediction model may be simplified by deleting this subset of variables.

a. Sums of Squares

To perform statistical significance tests all m of the b_j elements in matrix B are computed from Equation 66 and the error sum of squares SSE is computed from Equation 65. Then with some k ($k<m$) of the b_j set equal to zero all m-k of the remaining b_j elements in matrix B are recomputed along with a new and larger total error sum of squares, SST. This new total error is larger because the assumption that k of the b_j equal zero has changed them from the minimizing values previously computed. The difference between this larger total error sum of squares (computed from new minimizing values for those b_j not assumed to be zero) and the original error sum of squares (computed from minimizing values for all the b_j) is called the hypothesis sum of squares, SSH. This relationship can be symbolized by the right triangle shown in Figure 14.

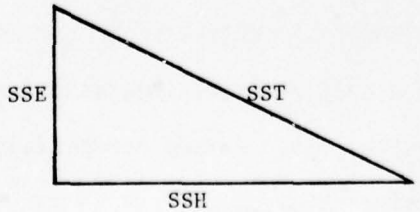


Figure 14. Error, Total and Hypothesis Sums of Squares

The relation of these sums of squares to other such sums is given in Appendix E. There, Figure 14 corresponds to the triangle TRS.

b. Variance Estimates

The number of degrees of freedom associated with SSE is the sample size n minus the number m of the b_j computed. Dividing SSE by n-m gives an estimate of the variance of the data about the regression function containing all m of the b_j. Likewise the number of degrees of freedom associated with SST is the sample size n minus the number m-k of the b_j computed, and dividing SST by n-(m-k) gives an estimate of the variance of the data about the regression function containing m-k of the b_j. This information is summarized in an analysis of variance table as follows.

Analysis of Covariance

Source	Sum of Squares	df	Mean Square
Hypothesis	SSH	k	SSH/k
Error	SSE	n-m	SSE/(n-m)
Total	SST	n-m+k	

If zero-valued population parameters actually do correspond to k of the b_j elements set equal to zero, then both SSE/(n-m) and SST/(n-m+k) estimate the same variance. In this case the hypothesis sum of squares SSH arises solely from sampling variation and dividing SSH by its k degrees of freedom yields another estimate of the population variance. As suggested by the perpendicular lines in Figure 14, SSH/k and SSE/(n-m) are independent estimates of variance. As previously noted the ratio of two independent estimates is associated with the F statistic.

c. Variance Ratio Test

This F ratio can be used to test statistically the hypothesis of zero values for the k population parameters corresponding to b_j elements set equal to zero in SSH. The testing procedure is as follows.

(1) Choose the acceptable level of risk - the probability of rejecting the hypothesis when it is in fact true.

(2) From a statistical F table for that level of risk select the tabulated F value for k degrees of freedom in the numerator and m-n degrees of freedom in the denominator.

(3) Obtain the computed F ratio from

$$F = \frac{SSH/k}{SSE/(n-m)}$$

(4) If this computed F exceeds the tabulated F reject the hypothesis of zero values for all k of the population parameters corresponding to the b_j elements set equal to zero in SSH.

(5) If the hypothesis of zero values for the k population parameters is rejected, the values obtained in B of Equation 66 can be taken as the best estimates with existing data.

4. TRANSFORMED GENERAL LINEAR HYPOTHESIS MODELS

If each measurement and associated covariate is replaced by its logarithm, Equation 64 becomes

$$\log y_i = \sum_{j=1}^{m} b_j \log x_{ji} + e_i \log 10$$

$$\log y_i = \sum_{j=1}^{m} \log x_{ji}^{b_j} + \log 10^{e_i} \tag{67}$$

$$\log y_i = \log \prod_{j=1}^{m} x_{ji}^{b_j} 10^{e_i}$$

$$y_i = \prod_{j=1}^{m} x_{ji}^{b_j} e_i' \qquad \text{where } e_i' = 10^{e_i}$$

This product results when input observations are subjected to a logarithmic transformation before using a general linear hypothesis procedure. It is the appropriate model to use for statistical prediction functions if the error terms are multiplicative, with 1 playing the same role that zero does in the additive case. Note from Equation 67 that $e_i' = 1$ when $e_i = 0$ (from $e_i' = 10^{e_i}$). This implies a log normal distribution function meaning that the logarithms of the errors are normally distributed.

SECTION VIII

CONCLUSIONS

Given a set of observations on a set of random variables an orderly data processing procedure would be precisely the sequence of operations given in the preceding sections of this report.

a. Compute univariate statistics giving some measure of average value, dispersion, skewness, and kurtosis for each of the random variables.

b. Select the probability density function associated with each variable by matching the computed univariate statistics with those tabulated for specific mathematical functions.

c. Conduct statistical t and F tests to determine if significant differences exist in the means and variances of random variables measured at different locations or under different test conditions.

d. Compute the bivariate correlations between each pair of random variables, arrange them in a correlation matrix, and then compute the multiple, marginal, conditional, and canonical correlations of particular interest. Perform a factor analysis on the correlation matrix to determine the structure of interrelations among the variables and how each variable may be expressed in terms of a smaller number of underlying factors.

e. Formulate mathematical models quantifying the precise relationship between any dependent variable and a particular set of independent variables or factors.

APPENDIX A

COEFFICIENT OF RANK CORRELATION

Given n bivariate observations let x_i, y_i represent respectively the rank of the ith observation of x among all x's and the ith observation of y among all y's. Then we have

$$\Sigma x_i = \Sigma y_i = n(n+1)/2$$

$$\Sigma x_i^2 = \Sigma y_i^2 = n(n+1)(2n+1)/6$$

$$\Sigma(x_i - y_i)^2 = \Sigma x_i^2 + \Sigma y_i^2 - 2\ \Sigma x_i\, y_i$$

$$\Sigma x_i\, y_i = [\Sigma x_i^2 + \Sigma y_i^2 - \Sigma(x_i - y_i)^2]/2$$

$$\Sigma x_i\, y_i = [2n(n+1)(2n+1) - 6\ \Sigma(x_i - y_i)^2]/12$$

$$r = \frac{(\Sigma\, x_i\ y_i/n) - (\Sigma\, x_i/n)(\Sigma\, y_i/n)}{\sqrt{(\Sigma x_i^2/n) - (\Sigma x_i/n)^2}\ \sqrt{(\Sigma y_i^2/n) - (\Sigma y_i/n)^2}}$$; if we let $d_i = x_i - y_i$

$$r = \frac{\{[2n(n+1)(2n+1) - 6\ d_i^2]/12n\} - [n(n+1)/2n]^2}{[n(n+1)(2n+1)/6n] - [n(n+1)/2n]^2}$$

$$r = \frac{[(4n^3 + 6n^2 + 2n - 6\Sigma d_i^2)/12n] - [(n^2 + 2n + 1)/4]}{[2n^3 + 3n^2 + n)/6n] - [(n^2 + 2n + 1)/4]}$$

$$r = \frac{4n^3 + 6n^2 + 2n - 6\Sigma\, d_i^2 - 3n^3 - 6n^2 - 3n}{4n^3 + 6n^2 + 2n - 3n^3 - 6n^2 - 3n}$$

$$r = \frac{n^3 - n - 6\ \Sigma\ d_i^2}{n^3 - n} = \frac{n(n^2 - 1) - 6\ \Sigma\ d_i^2}{n(n^2 - 1)} = 1 - \frac{6\ \Sigma\ d_i^2}{n(n^2 - 1)}$$

This is the coefficient of rank correlation.

PRECEDING PAGE BLANK-NOT FILMED

APPENDIX B

POINT BISERIAL CORRELATION

Given n bivariate observations x_i, y_i, the first of which is either one or zero according as to whether a given attribute is or is not present when the y measurement is taken. To indicate these two conditions zero and one subscripts are used with the measurement y and the sample size n in the following derivation

$$\Sigma x_i = \Sigma x_i^2 = n_1 \qquad \Sigma x_i\, y_i = \Sigma y_{1i}$$

$$r_p = \frac{(\Sigma x_i\, y_i/n) - (\Sigma x_i/n)(\Sigma y_i/n)}{\sqrt{(\Sigma x_i^2/n) - (\Sigma x_i/n)^2}\ \sqrt{(\Sigma y_i^2/n) - (\Sigma y_i/n)^2}}$$

$$r_p = \frac{(\Sigma y_{1i}/n) - (n_1)\ (\Sigma y_{0i} + \Sigma y_{1i})/n^2}{\sqrt{(n_1/n) - (n_1/n)^2}\ \sqrt{\overline{y^2} - \bar{y}^2}}$$

$$r_p = \frac{(n_1/n)\ (\bar{y}_1) - (n_0\, n_1/n^2)\bar{y}_0 - (n_1^2/n^2)\bar{y}_1}{\sqrt{(n_1/n)\ (1 - n_1/n)}\ \sqrt{\overline{y^2} - \bar{y}^2}}$$

$$r_p = \frac{(n_1/n)\ (1 - n_1/n)\bar{y}_1 - (n_0\, n_1/n^2)\bar{y}_0}{\sqrt{(n_1/n)\ (n_0/n)}\ \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{(n_0\, n_1/n^2)\ (\bar{y}_1 - \bar{y}_0)}{\sqrt{(n_0\, n_1/n^2)}\sqrt{\overline{y^2} - \bar{y}^2}}$$

$$r_p = \frac{\bar{y}_1 - \bar{y}_0}{s_y}\ \sqrt{\frac{n_0}{n}\ \frac{n_1}{n}} = \frac{\bar{y}_1 - \bar{y}_0}{s_y}\ \sqrt{p_0\, p_1}$$

This is the point biserial coefficient of correlation.

APPENDIX C

TETRACHORIC CORRELATION

Given n bivariate observations x_i, y_i each of which is either one or zero according to whether each of a pair of attributes is or is not present. The sample may be subdivided as follows

a = number of observations for which x = 0, y = 0

b = number of observations for which x = 1, y = 0

c = number of observations for which x = 0, y = 1

d = number of observations for which x = 1, y = 1

Then $n = a + b + c + d$

$$\Sigma x_i = \Sigma x_i^2 = b + d, \qquad \Sigma y_i = \Sigma y_i^2 = c + d \qquad \Sigma x_i\, y_i = d$$

$$r_\phi = \frac{\sum_{i=1}^{n} x_i y_i/n - \left(\sum_{i=1}^{n} x_i/n\right)\left(\sum_{i=1}^{n} y_i/n\right)}{\sqrt{\left(\sum_{i=1}^{n} x_i^2/n\right) - \left(\sum_{i=1}^{n} x_i/n\right)^2}\sqrt{\left(\sum_{i=1}^{n} y_i^2/n\right) - \left(\sum_{i=1}^{n} y_i/n\right)^2}}$$

$$r_\phi = \frac{\frac{d}{a+b+c+d} - \left(\frac{b+d}{a+b+c+d}\right)\left(\frac{c+d}{a+b+c+d}\right)}{\sqrt{\frac{b+d}{a+b+c+d} - \left(\frac{b+d}{a+b+c+d}\right)^2}\sqrt{\frac{c+d}{a+b+c+d} - \left(\frac{c+d}{a+b+c+d}\right)^2}}$$

$$r_\phi = \frac{(ad + bd + cd + d^2) - (bc + bd + cd + d^2)}{\sqrt{(a+b+c+d)(b+d) - (b+d)^2}\sqrt{(a+b+c+d)(c+d) - (c+d)^2}}$$

$$r_\phi = \frac{ad + bd + cd + d^2 - bc - bd - cd - d^2}{\sqrt{[(b+d) + (a+c)]\,(b+d) - (b+d)^2}\ \sqrt{[(a+b) + (c+d)]\,(c+d) - (c+d)^2}}$$

$$r_\phi = (ad - bc)/\sqrt{(a+c)(b+d)}\sqrt{(a+b)(c+d)}$$

This is the tetrachoric coefficient of correlation.

APPENDIX D

COEFFICIENT OF CONTINGENCY

If we let f_{ij} be the actual number of observations in row i and column j of a contingency table and f_{Oij} be the expected number of observations in the same position assuming no interdependence, then the quantity chi square (χ^2) is given by the sum

$$\chi^2 = \Sigma_i \Sigma_j \frac{(f_{ij} - f_{Oij})^2}{f_{Oij}} = \Sigma_i \Sigma_j \left[\frac{f_{ij}^2 + f_{Oij}^2 - 2 f_{ij} f_{Oij}}{f_{Oij}} \right]$$

$$\chi^2 = \Sigma_i \Sigma_j \left[\frac{f_{ij}^2}{f_{Oij}} + f_{Oij} - 2 f_{ij} \right] = \Sigma_i \Sigma_j \frac{f_{ij}^2}{f_{Oij}} + n - 2n$$

$$\chi^2 = \Sigma_i \Sigma_j \frac{f_{ij}^2}{f_{Oij}} - n$$

However $f_{Oij} = r_i c_j / n$ where r_i is the total number of observations in row i and c_j is the total number of observations in column j. Therefore

$$\chi^2 = \Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j / n} - n = \Sigma_i \Sigma_j \frac{n f_{ij}^2}{r_i c_j} - n$$

$$\chi^2 = n \left[\Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j} - 1 \right]$$

Using this expression for χ^2 in the definition of the coefficients of contingency gives the following result:

$$\sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{n\left[\Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j} - 1\right]}{n\left[\Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j} - 1\right] + n}}$$

$$\sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{n\left[\Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j} - 1\right]}{n\left[\Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j} - 1 + 1\right]}}$$

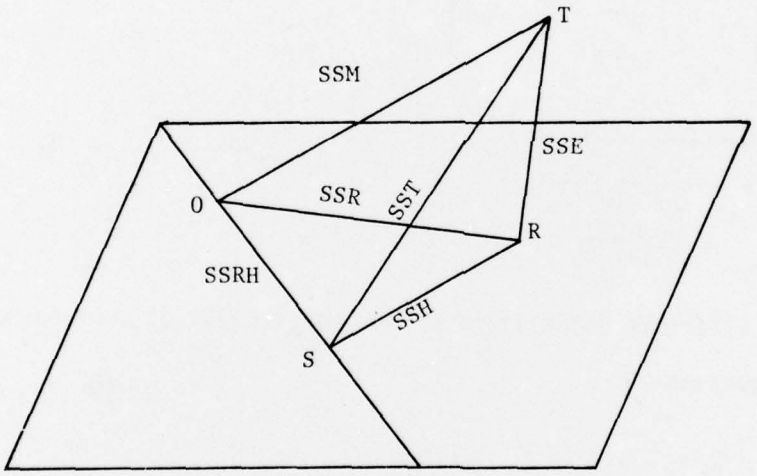
$$\sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{1 - \frac{1}{\Sigma_i \Sigma_j \frac{f_{ij}^2}{r_i c_j}}}$$

These are the alternate definitions of the coefficient of contingency as given in Equation 29.

APPENDIX E

SUMS OF SQUARES IN ANALYSIS OF VARIANCE

In mathematical models for statistical data, measurements are resolved into two components: explained by regression and unexplained error. The relationship among the sums of squares of these components are indicated in the figure below. Projections first onto the plane and then onto the line symbolize regression functions with a limited and then further reduced number of variables.



Vector OT - represents the measurement sum of squares SSM. Each measurement is squared and then all are summed.

Vector OR - the projection of OT on the plane, represents the regression sum of squares SSR. Each regression estimate is squared and then all are summed.

Vector OS - the projection of both OR and OT on the line, represents the hypothesis regression sum of squares for a reduced number of

variables SSRH. Each such reduced regression estimate is squared and then all are summed.

Vector SR - the difference between OR and OS, represents the hypothesis sum of squares SSH. The difference between each regression and reduced regression estimate is squared and then all are summed. In a simple one-way analysis of variance this is the among-groups sum of squares.

Vector RT - the difference between OT and OR, represents the error sum of squares SSE. The difference between each measurement and its regression estimate is squared and then all are summed. In a simple one-way analysis of variance this is the within-groups sum of squares.

Vector ST - the difference between OT and OS, and also the sum of SR and RT, represents the total sum of squares, SST. The difference between each measurement and its reduced regression estimate is squared and then all are summed. Alternatively SST=SSH+SSE, this relation being the reason for the name "total sum of squares".

As a numerical illustration of these sums of squares consider the following one-way analysis of variance of four groups of six observations each. In each group the observed measurement on the right of the equal sign is shown as the sum of a common value plus a group effect plus an error term.

Group 1	Group 2	Group 3	Group 4
80+60+2 = 142	80+0+11 = 91	80-20+14 = 74	80-40+16 = 56
80+60+2 = 142	80+0+ 9 = 89	80-20+12 = 72	80-40+14 = 54
80+60+1 = 141	80+0+ 2 = 82	80-20+ 1 = 61	80-40+ 1 = 41
80+60-1 = 139	80+0- 2 = 78	80-20- 1 = 59	80-40- 1 = 39
80+60-1 = 139	80+0- 9 = 71	80-20-13 = 47	80-40-15 = 25
80+60-3 = 137	80+0-11 = 69	80-20-13 = 47	80-40-15 = 25

Note that the error values sum to zero within groups and the group effects sum to zero across groups. The squares of these numbers are tabulated below:

6400	3600	4	20164	6400	0	121	8281	6400	400	196	5476	6400	1600	256	3136
6400	3600	4	20164	6400	0	81	7921	6400	400	144	5184	6400	1600	196	2916
6400	3600	1	19881	6400	0	4	6724	6400	400	1	3721	6400	1600	1	1681
6400	3600	1	19321	6400	0	4	6084	6400	400	1	3481	6400	1600	1	1521
6400	3600	1	19321	6400	0	81	5041	6400	400	169	2209	6400	1600	225	625
6400	3600	9	18769	6400	0	121	4761	6400	400	169	2209	6400	1600	225	625

Summing corresponding columns for each group gives

$$24\ (6400) + 6(3600 + 0 + 400 + 1600) + 2016 = 189{,}216$$

$$153{,}600 + 33{,}600 + 2016 = 189{,}216$$

SSH = 33,600 Among-groups or hypothesis sum of squares

SSE = 2,016 Within-groups or error sum of squares

SST = 33,600 + 2,016 Total sum of squares

SSRH = 153,600 = Reduced regression sum of squares

SSR = 153,600 + 33,600 = Regression sum of squares

SSM = 153,600 + 33,600 + 2016 = Measurement sum of squares

All sums of squares are thus obtainable from simple summations for this special case in which SSH is simply the sum of the squared group effects. However, in more general cases (unequal group sample sizes, for example), the value of the common term changes when group effects are assumed to be zero and the effect of this difference must also be a part of SSH.

BIBLIOGRAPHY

General

Dixon, Wilfrid J. and Massey, Frank J., Introduction to Statistical Analysis, 3rd Edition, McGraw Hill Book Company, Inc. New York, N. Y., 1969.

An exceptionally broad range of statistical analysis techniques for an introductory text are treated in this book. Statistical tables are more diversified and complete than found in similar texts. First author is also editor of the BMD statistical computer program manual cited below.

Guilford, Joy Paul, Fundamental Statistics in Psychology and Educuation, 4th edition, McGraw Hill Book Company, Inc. New York, N. Y., 1965.

Correlation analysis is more comprehensively treated than in books by mathematical statisticians. Numerical illustrations are used extensively in treating each topic.

Hahn, Gerald J. and Shapiro, Samuel S., Statistical Models in Engineering, John Wiley and Sons, Inc., New York, N. Y., 1967.

Non-normal (non-Gaussian) distributions are treated more comprehensively than is the case in more conventional statistics books. Modeling method developed is oriented toward systems analysis.

Specialized Books

Bendat, Julius S. and Piersol, Allan G., Random Data Analysis and Measurement Procedures,, John Wiley and Sons, Inc., New York, N.Y., 1971.

Randomly varying time series measurements are treated in this book which is oriented towards systems analysis.

Harman, Harry H., Modern Factor Analysis, 2nd edition, University of Chicago, 1967.

The factor analysis of correlation matrices is treated in this book.

Scheffé, Henry, The Analysis of Variance, John Wiley and Sons, Inc., New York, N. Y., 1959.

The mathematical modeling of statistical measurments is treated in this book.

Reference Books

Dixon, Wilfrid J., BMD: Biomedical Computer Programs, University of California, 1975.

Computer programs widely available at scientific computer facilities throughout the country are described in this manual. The title notwithstanding, the programs are standard statistical analysis methods that can be used for data from any source.

Johnson, Norman L. and Kotz, Samuel, Discrete Distributions, 1969, and Continuous Univariate Distributions, Vols. I and II, 1970, John Wiley and Sons, Inc., New York, N. Y.

Probability density functions of many kinds are defined and described in these volumes. Information about each includes historical origin, mathematical characteristics, important applications, and relationships to other probability densities. Also more recent volumes deal with multivariate continuous distributions.

Natrella, Mary Gibbons, Experimental Statistics, U.S. Government Printing Office, 1963.

Statistical tests and analysis methods for a broad range of statistical measurements are set forth along with a detailed description of the procedures involved in each case. The sections on experimental designs are widely cited.

☆ U. S. GOVERNMENT PRINTING OFFICE: 1977 — 757-001/254